

# BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006

Miguel Á. García-Cumbreras, Fernando Martínez-Santiago  
L. Alfonso Ureña-López and José M. Perea-Ortega

SINAI Research Group. Computer Science Department. University of Jaén. Spain  
{dofer,magc,laurena,jmperea}@ujaen.es

**Abstract.** This paper presents our first participation in the bilingual English-Spanish track at QA@CLEF 2006. The contribution is the Multilingual BRUJA system, a Question Answering (QA) system that works with questions in several languages and also collections in several languages.<sup>1</sup>

## 1 Introduction

A Question Answering (QA) system searches and shows the user an accurate and concise answer, given a free-form question, and using a large text data collection [1].

The use of Cross Language Information Retrieval Systems (CLIR) and their application to other general systems, such as Question Answering or Question Classification, are clearly on the increase.

This is the first participation for the SINAI group of the University of Jaén (Intelligent Systems of Information Access) at CLEF-QA. We have accomplished a bilingual task, from English to Spanish.

We present the BRUJA system (*Búsqueda de Respuestas University of JAén*; in English, Question Answering University of Jaén), a prototype of a complete multilingual QA system. Based on NLP tools, it works with questions in three languages, English, Spanish and French) and also with collections in these languages.

This paper describes the system architecture and some details of each module (Section 2) and the main experiments and the results obtained (Section 3). Conclusions and further research are presented in Section 4.

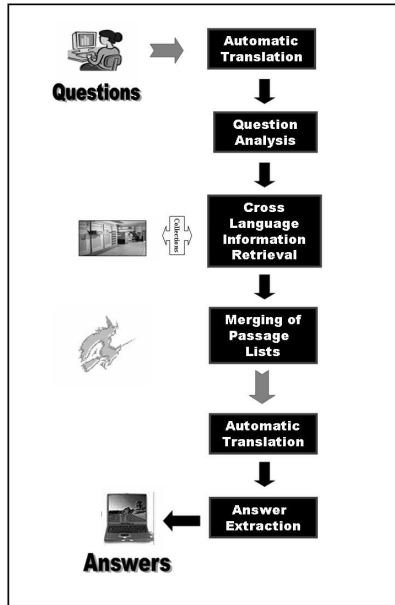
## 2 System Description

In this section our multilingual QA system is presented. The development of the BRUJA system is in its first phase, so some modules are not finished yet and others are still being tuned and corrected.

---

<sup>1</sup> This work has been supported by the Spanish Government with grant TIC2003-07158-C04-04

## 2.1 Overview



The BRUJA system is a prototype of a complete multilingual Question Answering system, which works with questions and collections in Spanish, English or French.

In short, when a new question arises it is translated to the other languages and the original questions and its translations are launched over their collection index. Then it is necessary to merge the monolingual lists of relevant documents or passages and one multilingual relevant list is returned, as in a usual Cross-Language Information Retrieval system (CLIR). The figure shows the architecture of the BRUJA system.

In the following sections each module in detail is described.

## 2.2 Translation and Question Analysis

When a new question arises its language is detected, and, if it is different from English, it is translated into English. This process uses SINTRAM (SINai TRAnslation Module)<sup>2</sup>, our machine translation system that works with different online machine translators and implements some heuristics.

This is because our preprocessing methods work with English questions, in order to improve the result of this phase. Preprocessing is made using the GATE architecture<sup>3</sup>.

After that, we run our Question Classification (QC) subsystem, which classifies the question under a general class (ABBR, DESC, ENTY, HUM, LOC or NUM)[2].

As data of this first phase we obtain relevant features, such as the focus of the questions, and the general class of the question.

## 2.3 Relevant passages versus relevant documents

In the Information Retrieval subsystem, we preprocess the collections first, using also GATE, and then we index these collections. In order to improve the results of this step, we index documents using LEMUR<sup>4</sup> and passages using the IR-n system developed at the University of Alicante.

<sup>2</sup> <http://sinai.ujaen.es>

<sup>3</sup> <http://gate.ac.uk>

<sup>4</sup> <http://www.lemurproject.org>

After that the results are combined with a simple voting system and score of each individual relevant docid is summed.

## 2.4 Passage selection improvement

Before extracting the answers we try to improve the passage selection. By default only the 10 first passages for each question are taken. The idea is to reduce the time consumption of the general system.

In this phase we apply some heuristics that depends on the question class. For instance, if the class of a question is LOC we expect the answer to be a location. In this case we take the ten first passages that contain any location applying Named Entity Recognition methods.

## 2.5 Answers extraction

The last stage of the QA system is answer extraction. This module takes the list of relevant passages and extracts and scores the possible answers for each question.

In the first step of this stage some patterns and heuristics are applied. The second step of the subsystem currently in progress, will be based on logic and machine learning.

In order to check the first prototype we have only tried to answer factual questions and unders some patterns for definitional questions.

For the factual ones some rules have been implemented based on the question class. For instance, for HUM questions Person Entities have been identified and extracted from the first ten relevant passages; for LOC questions Location Entities, and so on...

In order to score the final answers a simple method has been applied. If the keywords of the question appear in the passage, the score of the answers increases.

For the definitional questions some patterns are applied, for instance:

- |  |  |
|--|--|
| – Question:  | – Some patterns:   |
| <ul style="list-style-type: none"><li>• What is <i>Linux</i>? (Qué es <i>Linux</i>?)</li></ul> | <ul style="list-style-type: none"><li>• Linux is DEF</li><li>• DEF, Linux</li><li>• Linux, DEF</li></ul> |

For the answer file, scores below 0.5 are rated as incorrect answers and NIL is written.

## 3 Results

This section describes the result obtained with a simple run sent and the evaluation. The proposed system was applied to the set of 200 questions, although only factual and some definitional questions have been used. Table 1 shows the results for our run.

Right	39
Inexact	5
Unsupported	8
Wrong	138
Correct answer string NIL	18
Overall accuracy	20.53%
Accuracy over Factoid questions	17.12%
Accuracy over Definition questions	33.33%
Accuracy over Temporally restricted Factoid questions	0.00%
overall Confidence Weighted Score (CWS)	0.16384

**Table 1.** General results for the bilingual English-Spanish run

The results show that the answer extraction module does not work properly, and only a low percentage of factual questions has good answers.

For Factoid questions, a simple manual analysis of the experiment explained some of results obtained.

- Some questions do not have real relevant passages (the focus words have not been identified well).
- In the cases where relevant passages contain the possible answer the Named Entity Recognition system fails.

For Definition questions, the same manual analysis explained other results. The main one is that there are a lot of patterns, and the use of only some of them is not enough.

## 4 Conclusions and further research

For our first participation in QA@CLEF track we proposed a prototype of a multilingual QA system that works with English and Spanish questions, to search Spanish relevant documents.

For this prototype the answer extraction was in its first phase, so the results are obviously not good. However based on these experiments the multilingual system will be assessed in several aspects.

Further research includes completing and tuning each module. The next important task is to evaluate the experiments made this year and develop the answer extraction module based on logic and machine learning.

## References

1. Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Petya Osenova, Anselmo Peas, Valentin Jijkoun, Bogdan Sacaleanu, Paulo Rocha, Richard Sutcliffe: Overview of the CLEF 2006 Multilingual Question Answering Track. In Proceedings of the Cross Language Evaluation Forum (CLEF 2006), 2006.
2. Miguel Á. García Cumbreñas and L.A. Ureña: BRUJA: Question Classification for Spanish. Using Machine Translation and an English Classifier. Conference of the European Chapter of the Association for Computational Linguistics, 2006.