BRUJA System. The University of Jaén at the Spanish task of CLEFQA 2006

Miguel A. García-Cumbreras, L.A. Ureña-López
Fernando Martínez Santiago, Jose M. Perea-Ortega
University of Jaén
{magc,laurena,dofer,jmperea}@ujaen.es

Abstract

This paper presents our first participation in the bilingual English-Spanish track at CLEF QA 2006. The Multilingual BRUJA system is presented, a Question Answering (QA) system that works with questions in several languages and also collections in several languages. The BRUJA system is currently in its first phase of develop, so we have only run one official experiment with questions into English and the collection into Spanish. The results obtained shown that the prototype and its answer extraction phase, have to be finished and improved. An overall accuracy of 20.53% in not a good result and the system is in progress.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Algorithms, Languages, Performance, Experimentation

Keywords

Information Retrieval, Question Answering for Spanish, Named Entity Recognition, Natural Language Processing

1 Introduction

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

The use of Cross Language Information Retrieval Systems (CLIR) is growing, and also the application of these ones into other general systems, such as Question Answering or Question Classification.

A CLIR system is an Information Retrieval System that works with collections in several languages, and extract relevant documents or passages [2].

The Cross-Language Evaluation Forum (CLEF) includes a multilingual forum to evaluate Question Answering Systems that works with several languages[3].

This is the first participation for the SINAI research group at the CLEF-QA task. We have accomplished a bilingual task, from English to Spanish.

The goal of this bilingual task is to answer a set of questions, where there is a questions language (English) and a different collection language (Spanish), so it is necessary to translate the set of questions.

We present the BRUJA system (in Spanish, Búsqueda de Respuestas University of JAén), a prototype of a complete multilingual QA system, based on NLP tools [4], that works with questions in several languages (actually three languages, English, Spanish and French) and also with collections in several languages (the same three).

We have combine different modules, in order to evaluate the system in different points, that are explained in the following sections.

Next Section describes the system architecture and some details of each module. In Section 3 we explain the main experiments and the results. Finally, conclusions and further works 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 tuned and corrected.

2.1 Overview

The BRUJA system is a prototype of a complete multilingual Question Answering system, that works with questions into Spanish, English or French, and the collection datasets are also into these three languages.

Basically, when a new question arise it is translated to the other languages and the original questions and its translations are launched over its collection index. Then it is necessary to merge the monolingual lists of relevant documents or passages and one multilingual relevant list is returned, like an usual Cross Language Information Retrieval system (CLIR).

In some steps we use English as the pivot language, and we apply online machine translators if it is necessary.

In the figure 1 we can see the architecture of the BRUJA system.

In the following sections we describe each module in detail.

2.2 Translation and Question Analysis

This is the first phase of the QA system.

When a new question arise we detect its language, and if it is different from English we translated it to English. We use SYMTRAM (our Machine Translation system that works with different online machine translators and implements some heuristics).

We do that because our preprocessing methods work with English questions, in order to improve the result of this phase. We make the preprocessing phase using the GATE architecture (REF).

For the main process and for the next one, the Question Classification, some lexical, syntatic and semantic features and the keywords are extracted.

After that, we run our Question Classification (QC) subsystem, that classify the question in a general class (ABBR, DESC, ENTY, HUM, LOC or NUM). Other experiments give us a high confidence in its module, with results around a 90% of F-score [1]. Our Question Classifier is based on machine learning, automatic online translators and different language features. It works with English collections and English monolingual questions or bilingual pairs (Spanish to English or French to English).

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

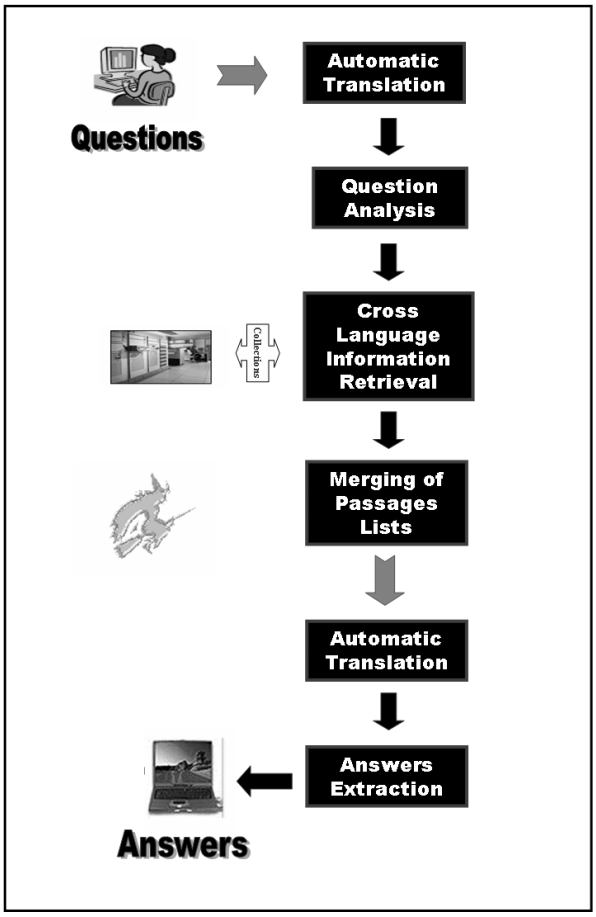

Figure 1: BRUJA system architecture

2.3 Relevant passages versus relevant documents

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

After that we combine the results with a simple voting system. We sum the score of each individual relevant docid.

The idea of the multilingual system is that it would have to join the lists of relevant passages from the different languages, to obtain one multilingual list. Because the merging module has not finished yet we have only worked with the Spanish list.

2.4 Passage selection improvement

Before to extract the answers we try to improve the passage selection. By default we take only the 10 first passages for each question. The idea is to decrease the time consumption of the general system.

In this phase we apply some heuristics that depends of the question class.

For instance if the class of a question is LOC we expect that the answer is a location. In this case we take the ten first passages that contains any location. To do that we apply some Named Entity Recognition methods.

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

2.5 Answers extraction

The final step of the QA system is the answer extraction. This module takes the list of relevant passages and extract and score the possible answers for each question.

In the first step some patterns and heuristics are applied. The second step of the subsystem will be based on logic and machine learning, but it is under development.

In order to check the first prototype we have only tried to answer factual questions and 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.

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 increase.

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

- Question:
  - What is Linux? (Qué es Linux?, in Spanish)
- Some patterns:
  - Linux is DEF
  - DEF, Linux
  - Linux, DEF

Finally, for the answers file, if the score is below 0.5, we consider that it is an incorrect answer, and NIL is written.

3 Results

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

Table 1 shows the results for our run.

The results shown that the answer extraction module don't work properly, and only a low percent of factual questions have good answers.

For Factoid questions a simple manual analysis of the experiment gave us some reasons about the results obtained.

- Some questions have not real relevant passages. This happen, for instance, when the focus words have not identified well, and the others keywords appear in the relevant passage.

- In the cases where relevant passages contains the possible answer we count on the goodness of the Named Entity Recognition system, and sometimes it fails or didn't recognize the entities.

For Definition questions the same manual analysis give us some reasons. 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 Future work

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 only the answers extraction was in its first phase, so the results are obviously not good, but with the experimentation made we will evaluate the multilingual system in different points.

As future work we will finish 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.

5 Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant TIC2003-07158-C04-04.

References

[1] Miguel Á. García Cumbreras and L.A. Ureña. Bruja: Question classification for spanish. using machine translation and an english classifier. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[2] Gregory Grefenstette, editor. *Cross-Language Information Retrieval*, volume 1. Kluwer academic publishers, Boston, USA, 1998.

[3] J. Herrera, A. Peñas, and F. Verdejo. Question answering pilot task at clef 2004. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Advances in Cross-Language Information Retrieval, CLEF 2004*, Lecture Notes in Computer Science, pages 445–45, 2004.

[4] S. Roger, S. Ferrández, A. Ferrández, J. Peral, F. Llopis, A. Aguilar, and D. Tomás. Aliqan, spanish qa system at clef-2005. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Advances in Cross-Language Information Retrieval, CLEF 2005*, Lecture Notes in Computer Science, 2005.