Does English helps Question Answering in Spanish?

José L. Vicedo, Maximiliano Saiz, and Rubén Izquierdo

Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
{vicedo,max,rib1}@dlsi.ua.es

Abstract. This paper describes the architecture, operation and results obtained with the Question Answering prototype for Spanish developed in the Department of Language Processing and Information Systems at the University of Alicante for CLEF-2004 Spanish monolingual QA evaluation task. Our system is based on the prototype developed for CLEF-2003 Spanish monolingual task [3]. This system has been enhanced mainly to use documents in different languages to obtain evidences for supporting and complementing CLEF Spanish corpora. Particularly, the experiments described are intended to study how to use English Web documents to support monolingual Spanish QA.

1 Introduction

Cross-Language Evaluation Forum Campaigns[1] (CLEF) are characterized for fostering investigation in multilingual information access systems from the perspective of European languages integration. Particularly last year, CLEF organized the *first Multiple Language Question Answering task (QA@CLEF-2003)* guided to the evaluation of QA systems in several languages. This evaluation was very important since it fostered the development of a series of resources for the development and evaluation of QA systems from a multilingual perspective.

This campaign, the *QA@CLEF-2004* proposed new difficulties and therefore, the characteristics of this evaluation changed significantly. Participants were provided with document collections and question sets in seven European languages: Spanish, Portuguese, Italian, Dutch, German, French and English.

Participants had to choose a language for questions and another for the target document collection. This way the evaluation proposed from monolingual tasks (when question and document languages were the same) to different combinations of bilingual QA (when selected languages were different).

For each language, the organisation provided 200 questions requiring factual or definition answers whose answer was not guaranteed to occur in the document collection. Systems should return only one response per question.

Our participation was restricted to the Spanish monolingual task. The novelty in the experiments developed was the use of documents in languages different

---

[1] http://clef-qa.itc.it/

from Spanish in order to obtain evidences for supporting answers obtained from CLEF Spanish corpora. Particularly, we performed monolingual task from two different perspectives: (1) using Web Spanish documents and (2) using English Web documents to support monolingual Spanish QA. This way we could be able to investigate on using English (or by extent, other languages) documents to support monolingual QA.

This paper is organised as follows: Section 2 describes the main characteristics of our QA system. Afterwards, we present and analyse the results obtained at QA@CLEF-2004 Spanish monolingual task. Finally, we extract initial conclusions and discuss directions for future work.

2 System Description

Our system is based on the QA system described in [3] where two main enhancements have been added: (1) the inclusion of a dictionary-based NE tagger and (2) the possibility of using Web documents in other languages to support monolingual Spanish QA.

As this system is described in detail in [3] we only present here their main characteristics and enhance the new modules included. Our system is organized into the following main modules:

1. Question analysis.
2. Passage retrieval.
3. Answer extraction.

Question analysis processes questions formulated to the system in order to detect and extract the useful information they contain. *Passage retrieval* module retrieves relevant passages from the Spanish EFE document collection and also from the Internet in the selected language (Spanish or English). Finally, the *answer selection* module processes relevant passages in order to locate and extract the final answer. Figure 1 shows system architecture.

2.1 Question Analysis

Question analysis module carries out two processes: *answer type classification* and *keyword selection*. The former detects the type of information that the question expects as answer (a date, a quantity, etc) and the latter selects those question terms (*keywords*) that will allow locating the documents that are likely to contain the answer. These processes are performed by using a manually developed set of lexical patterns. Answer types have increased and now the system currently copes with seven possible answer types: NUMBER, DATE, LOCATION, PERSON, ORGANIZATION, DEFINITION and OTHER.
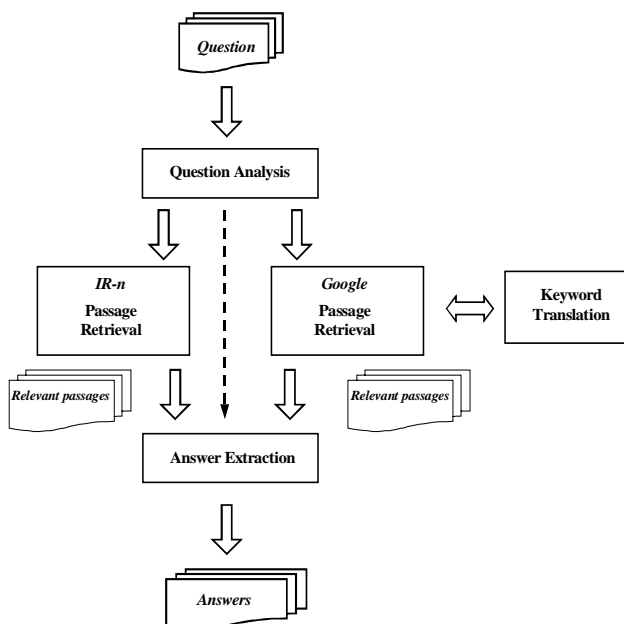
Question

Question Analysis

IR-n
Passage
Retrieval

Google
Passage
Retrieval

Keyword
Translation

Relevant passages

Relevant passages

Answer Extraction

Answers

Fig. 1. System architecture

2.2 Passage Retrieval

Passage retrieval stage is accomplished in parallel using two different search engines: IR-n [2] and Google[2]. IR-n system performs passage retrieval over the entire Spanish EFE document collection. In this case, keywords detected at question analysis stage are processed using MACO Spanish lemmatiser [1] and their corresponding lemmas are used for retrieving the 50 most relevant passages from the EFE document database.

In parallel, the same keyword list (without being lemmatised) is translated to the language the system is required to use for Web search (in this case Spanish or English) and posed to Google Internet search engine. The system selects the 50 best short summaries returned in Google main retrieval pages. Keywords have been translated by using SysTran[3] online translation services.

2.3 Answer Extraction

This module processes in parallel both sets of passages selected at passage retrieval stage (IR-n and Google) in order to detect and extract the most probable answer to the query. This process involves: (1) selecting and evaluating candidate answer from CLEF Spanish document collection and (2) adding web evidence

---

[2] http://www.google.com/
[3] http://www.systransoft.com/

to candidate list obtained from the Spanish collection. This process is explained in detail in [3]

3   Results

We submitted two runs. First run (*aliv041eses*) was obtained applying the system described above and using Spanish Web retrieved documents while second run performed QA process by activating English Web retrieval (*aliv042eses*). Table 1 shows the results obtained for each run.

Table 1. Spanish monolingual task results

Run	Accuracy (%)		
	Factoid	**Definition**	**Overall**
aliv041eses	30.56	40.00	31.50
aliv042eses	31.11	45.00	32.50

Result analysis shows that evidence obtained through English Internet documents (*aliv042eses*) performs better than using Spanish Web documents for this purpose (*aliv041eses*). Nevertheless, performance differences are near insignificant (32.5% − 31.5%). These results contradicted our initial hypotheses since we thought that English web documents would probably help more significantly Spanish monolingual QA. After a shallow error analysis we detected several translation problems that seriously affected the process of English Web document processing:

- *Keyword translation*. The lack of context when translating question keywords produces non-adequate translations. This implies sometimes retrieving English documents that have no semantic relation with the original Spanish question.
- *Proper noun translation*. Proper noun translation is an unresolved problem. Usually, proper nouns referring to people or companies have no translation (eg. Bill Clinton). On the other hand, names of countries (*España vs.* Spain) or cities (*Londres vs.* London) differ depending on the language.
- *Abbreviation translation*. Abbreviations usually refer to language-dependent expressions. From this point of view, if we want to correctly translate abbreviations and acronyms we need to know the whole expression or terms they refer to in the original language.
- *Titles translation*. Titles, such as names of books or films should not be translated. The basic problem here resides in detecting these expressions to be excluded from translation processes.

All these translation problems affect passage retrieval and answer extraction stages. First, an incorrect translation of content words in questions supposes the

retrieval of useless documents that do not support the original question. And second, it makes impossible to take advantage of evidences in other languages for supporting candidate answer selection if proper nouns, abbreviations an titles are not correctly translated.

4 Future Work

This work is a first attempt to perform monolingual QA in Spanish by using evidences obtained form corpora in different languages, in this case, English.

As we have previously seen, using corpora in other languages to support monolingual QA is possible and worthwhile if we are able to solve correctly the translation problems described before. Consequently, main line of future work to be investigated will be directed to adopt translation techniques that minimize the currently detected errors.

Moreover we argue that surely, this problem is the main bottleneck towards the main long-term objective of developing a whole system capable of performing multilingual question answering.

5 Acknowledgements

References

1. J. Atserias, J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text. In *Proceedings of First International Conference on Language Resources and Evaluation. LREC'98*, pages 1267–1272, Granada, Spain, 1998.
2. F. Llopis, J.L. Vicedo, and A. Ferrández. IR-n system, a passage retrieval systema at CLEF 2001. In *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, Darmstadt, Germany, 2001. Springer-Verlag.
3. J.L. Vicedo, R. Izquierdo, F. Llopis, and R. Muñoz. Question Answering in Spanish. In *Workshop of Cross-Language Evaluation Forum (CLEF 2003)*, Lecture notes in Computer Science, Thondheim, Norway, 2003. Springer-Verlag.