Question Answering in Spanish

José L. Vicedo, Ruben Izquierdo, Fernando Llopis, and Rafael Muñoz

Departamento de Lenguajes y Sistemas Informáticos,
University of Alicante, Spain
{vicedo,rib1,llopis,rafael}@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 the CLEF 2003 Spanish monolingual QA evaluation task. Our system has been fully developed from scratch and it combines shallow natural language processing tools with statistical data redundancy techniques. The system is able to perform QA tasks independently from static corpora or from Web documents. Moreover, the World Wide Web can be used as an external resource to obtain evidence to support and complement the CLEF Spanish corpora.

1 Introduction

Open domain QA systems are defined as tools capable of extracting the answer to user queries directly from unrestricted domain documents. Investigation in question answering has been traditionally focussed to English language and mainly fostered by Text REtrieval Conference (TREC[1]) evaluations. However, the development of QA systems for languages other than English was considered by the QA Roadmap Committee as one of the main lines for future investigations in this field [1]. In particular, it recommended that systems should be developed that perform QA from sources of information written in different languages.

As result of this interest, the Cross-Language Evaluation Forum[2] (CLEF 2003), has introduced a new task (*Multiple Language Question Answering*) for the evaluation of QA systems in several languages. This evaluation offers several subtasks: monolingual Spanish, Italian and Dutch QA and bilingual QA. The bilingual subtask is designed to measure system performance when searching answers in a collection of English texts to questions posed in Spanish, Italian, Dutch, German or French.

The main characteristics of this first evaluation are similar to those proposed in past TREC Conferences. For each subtask, the organisation provided 200 questions requiring short, factual answers whose answer is not guaranteed to occur in the document collection. Systems should return up to three responses

[1] http://trec.nist.gov/
[2] http://clef-qa.itc.it/

C. Peters et al. (Eds.): CLEF 2003, LNCS 3237, pp. 541–548, 2004.

per question, and answers should be ordered by confidence. Responses have to be associated with the document in which they are found. A response can be either a [*answer-string, document-identifier*] pair or the string "NIL" when the system does not find a correct answer in the document collection. The "NIL" string is considered correct if there is no answer known to exist in the document collection; otherwise it is judged as incorrect. Two different kinds of answers are accepted: the exact answer or a 50 bytes long string that should contain the exact answer.

Our participation has been restricted to the Spanish monolingual task in the category of exact answers. Although we have experience in past TREC competitions [2, 3, 4], we decided to build a new system mainly due to the big differences between English and Spanish languages. Moreover, we designed a very simple approach (1 person month) that will facilitate later error analysis and will allow the detection of those basic language-dependent features that make Spanish QA different from English QA.

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

2 System Description

Our system is organized in the three main modules of a general QA system architecture:

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

Question analysis is the first stage in the QA process. This module processes questions input to the system in order to detect and extract the useful information contained. This information is represented in a form that is easily processible by the remaining modules. The *Passage retrieval* module performs a first selection of relevant passages. This process is accomplished in parallel retrieving relevant passages from the Spanish EFE document collection and the Spanish pages in the World Wide Web. 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

The question analysis module carries out two main 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 make it possible to locate those documents that are likely to contain the answer.

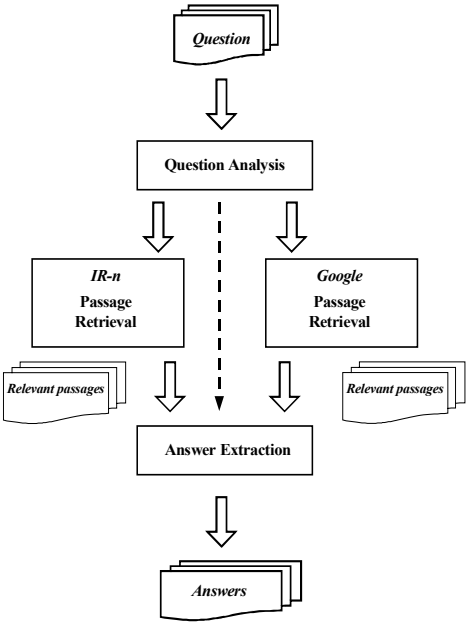

Fig. 1. System architecture

These processes are performed using a simple manually developed set of lexical patterns. Each pattern is associated with its corresponding expected answer type. Once a pattern matches the question posed to the system, this process returns both the list of keywords associated with the question and the type of the expected answer associated with the matched pattern. As our system lacks a named-entity tagger, it currently only copes with three possible answer types: NUMBER, DATE and OTHER. Figure 2 shows examples of the patterns and the output generated at the question analysis stage for test questions 002, 006 and 103.

2.2 Passage Retrieval

The passage retrieval stage is accomplished in parallel using two different search engines: IR-n [5] and Google[3].

IR-n is a passage retrieval system that uses groups of contiguous sentences as units of information. From the QA perspective, this passage extraction model allows us to benefit from the advantages of discourse-based passage retrieval models since self-contained information units of text, such as sentences, are used for building the passages. First, the IR-n system performs passage retrieval over the entire Spanish EFE document collection. In this case, keywords detected at

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

the question analysis stage are processed using the MACO Spanish lemmatiser [6] and their corresponding lemmas are used to retrieve the 50 most relevant passages from the EFE document database. These passages are made up of text snippets of 2 sentences. Second, the same keyword list (without lemmatisation) is input to the Google Internet search engine. For efficiency, relevant documents are not downloaded but the system just selects the 50 best short summaries returned in the main Google result pages. Figure 3 shows examples of retrieved passages for question 103. In this example, question keywords found in relevant passages are underlined.

Question 002	**¿Qué país invadió Kuwait en 1990?**
Pattern	`(qué|Qué)\s+([a-z|áéíóúñ]+)`
Answer type	`OTHER`
Keywords	`país invadió Kuwait 1990`
Lemmas	`país invadir Kuwait 1990`
Question 006	**¿Cuándo decidió Naciones Unidas imponer el embargo sobre Irak?**
Pattern	`(cuándo|Cuándo)\s+`
Answer type	`DATE`
Keywords	`decidió Naciones Unidas imponer embargo Irak`
Lemmas	`decidir Naciones Unidas imponer embargo Irak`
Question 103	**¿De cuántas muertes son responsables los Jemeres Rojos?**
Pattern	`(Cuántos|cuántos|Cuántas|cuántas)\s+([a-z|áéíóúñ]+)`
Answer type	`NUMBER`
Keywords	`muertes responsables Jemeres Rojos`
Lemmas	`muerte responsable Jemeres Rojos`

Fig. 2. Question analysis example

2.3 Answer Extraction

This module processes in parallel both sets of passages selected at the passage retrieval stage (IR-n and Google) in order to detect and extract the three most probable answers to the query. The processes involved at this stage are the following:

1. *Relevant Sentence Selection.* Sentences in relevant passages are selected and scored.
   (a) Passages are split into sentences.
   (b) Each sentence is scored according to the number of question keywords they contain. Keywords appearing twice or more are only added once. This value (*sentence_score*) measures the similarity between each relevant sentence and the question.
   (c) Sentences that do not contain any keyword are discarded (*sentence_score = 0*).

Question 103 **¿De cuántas muertes son responsables los Jemeres Rojos?**

First retrieved passage from EFE Collection:

<DOCNO> EFE19940913-06889
... explotan los Jemeres Rojos, quienes no les preocupa que sus ideas no sean respetadas por la comunidad internacional, que los acusa de ser los responsables de la muerte de más de **un millón** de camboyanos durante el genocidio de 1975 1978.

First retrieved passage from the World Wide Web:

<DOCNO> 1 Gooogle
Los Jemeres Rojos fueron responsables de más de **un millón** de muertes, mataron al menos a **20.000** presos políticos y torturaron a **cientos de miles** de personas.

Fig. 3. Passages retrieved for question 103

2. *Candidate Answer Selection.* Candidate answers are selected from relevant sentences.
   (a) Relevant sentences are tagged using the MACO lemmatizer.
   (b) Quantities, dates and proper noun sequences are detected and are merged into single expressions.
   (c) Every term or merged expression in relevant sentences is considered a candidate answer.
   (d) Candidate answers are filtered. This process gets rid of those candidates that start or finish with a stopword or contain a question keyword.
   (e) From the remaining candidate set, only those whose semantic type matches the expected answer type are selected. When the expected answer type is OTHER, only proper noun phrases are selected as final candidate answers. Figure 3 shows (in boldface) the selected answer candidates for question 103.
3. *Candidate Answer Combination.* Each answer candidate is assigned a score that measures its probability of being the correct answer (*answer_frequency*). As the same candidate answer can probably be found in different relevant sentences, the candidate answer set may contain repeated elements. Our system exploits this fact by relating candidate redundancy with answer correctness as follows:
   (a) Repeated candidate answers are merged into a single expression that is scored according to its frequency in the candidate answer set.
   (b) Shorter expressions are preferred as answer to longer ones. This way, terms in long candidates that appear themselves as answer candidates boost shorter candidate answer scores by adding long candidate scores to the frequency value obtained by shorter ones.

Table 1. Spanish monolingual task results

	Strict		Lenient	
Run	**MRR**	**Correct (%)**	**MRR**	**Correct (%)**
alicex031ms	0,3075	40,0	0,3208	43,5
alicex032ms	0,2966	35,0	0,3175	38,5

4. *Web Evidence Addition.* At this point the system has two lists of candidate answers: one obtained from the EFE document set and another from available Spanish web documents. Next, both candidate answer lists are merged. This process consists of increasing the answer frequency of EFE list candidates by adding their corresponding frequency values from the web list. In this way, candidates appearing only in the web list are discarded.
5. *Final Answer Selection.* Answer candidates from previous steps are given a final score (*answer_score*) that measures two circumstances: (1) their redundancy through the answer extraction process (*answer_frequency*) and (2) the context in which they have been found (*sentence_score*). As the same candidate answer may be found in different contexts, an answer will maintain the maximum score for all the contexts they appear in. The final answer score is computed as follows:

$$answer\_score = sentence\_score \cdot answer\_frequency \tag{1}$$

Answers are then ranked accordingly to their answer score and the first three answers are selected for presentation. Among the candidate answers for question 103 (example in Figure 3), the system selects "*un millón*" (one million) as the top ranked answer.

3 Results

We submitted two runs for the exact answer category. The first run (*alicex031ms*) was obtained applying the whole system as described above, while second run performed the QA process without activating Web retrieval (*alicex032ms*). Table 1 shows the results obtained for each run.

The result analysis may not be as conclusive as we would like, mainly due to the simplicity of our approach. Besides, the lack of the correct answers for test questions at this moment does not allow us to perform a correct error analysis. In any case, the results obtained show that using the World Wide Web as an external resource increases the percentage of correct answers retrieved by five percentage points. This fact confirms that the performance of QA systems for languages other than English can also benefit from this resource.

4 Future Work

This work has to be seen as a first and simple attempt to perform QA in Spanish. Consequently, there are several areas for future work to be investigated. Among them, we can indicate the following:

- Question analysis. Since the same question can be formulated in very diverse forms (interrogative, affirmative, using different words and structures,...), we need to study aspects such as recognizing equivalent questions regardless of the speech act or of the words, syntactic and semantic inter-relations or idiomatic forms employed.
- Answer taxonomy. An important part in the process of question interpretation resides in the system's ability to relate questions with the characteristics of their respective answers. Consequently, we need to develop a broad answer taxonomy that enables multilingual answer type classification. We expect to do this using the EuroWordNet[4] semantic net structure.
- Passage retrieval. An enhanced question analysis will improve passage retrieval performance by including question expansion techniques that make it possible to retrieve passages including relevant information expressed with terms that are different (but equivalent) to those used for question formulation.
- Answer extraction. Using a broad answer taxonomy involves using tools capable of identifying the entity that a question expects as answer. Therefore we need to integrate named-entity tagging capabilities that make it possible to narrow down the number of candidates to be considered as answers to a question.

Even though all these issues need to be investigated, it is important to note that this research needs to be developed from a multilingual perspective. Future investigations must address language-dependent and language-independent module detection in combination with the main long-term objective of developing a complete system capable of performing multilingual question answering.

Acknowledgements

This work has been partially supported by the Spanish Government (CICYT) with grant TIC2003-07158-C04-01.

References

1. Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E., Weishedel, R.: Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc (2000)

[4] http://www.dcs.shef.ac.uk/nlp/funded/eurowordnet.html

2. Vicedo, J., Ferrández, A.: A semantic approach to Question Answering systems. In: Ninth Text REtrieval Conference. Volume 500-249 of NIST Special Publication., Gaithersburg, USA, National Institute of Standards and Technology (2000) 511–516
3. Vicedo, J., Ferrández, A., Llopis, F.: University of Alicante at TREC-10. In: Tenth Text REtrieval Conference. Volume 500-250 of NIST Special Publication., Gaithersburg, USA, National Institute of Standards and Technology (2001)
4. Vicedo, J., Llopis, F., Ferrández, A.: University of Alicante Experiments at TREC-2002. In: Eleventh Text REtrieval Conference. Volume 500-251 of NIST Special Publication., Gaithersburg, USA, National Institute of Standards and Technology (2002)
5. Llopis, F., Vicedo, J., Ferrández, A.: 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, Springer-Verlag (2001)
6. Atserias, J., Carmona, J., Castellón, I., Cervell, S., Civit, M., Màrquez, L., Martí, M., Padró, L., Placer, R., Rodríguez, H., Taulé, M., Turmo, J.: Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text. In: Proceedings of First International Conference on Language Resources and Evaluation. LREC'98, Granada, Spain (1998) 1267–1272