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 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. This system is able to perform QA tasks independently from static corpora or from Web documents. Moreover, World Wide Web can be used as external resource to obtain evidences for supporting and complementing 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 mainly fostered by TREC[1] evaluations. However, the developing of QA systems for other languages than English was considered by the QA Roadmap Committee as one of the main lines of future investigations in this field [2]. Moreover, it considered essential obtaining systems 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 organised a new task (*Multiple Language Question Answering*) guided to the evaluation of QA systems in several languages. This evaluation proposes several subtasks: monolingual Spanish, Italian and Dutch QA and bilingual QA. The bilingual subtask is designed to measure systems ability in finding answers in a collection of English texts, when questions are 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 provides 200 questions requiring short, factual answers whose answer is not guaranteed to occur in the document collection. Systems should return up to three responses per question, and answers should be ordered by confidence. Responses have to

[1] Text REtrieval Conference
[2] http://clef-qa.itc.it/

be associated to the document they are found in. A response can be either a [*answer-string, docid*] pair or the string "NIL" when the systems do 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 [4–6], 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 detecting those basic language-dependent characteristics 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 QA system is structured into 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 QA process. This module processes questions formulated to the system in order to detect and extract the useful information they contain. This information is represented in a form that allows to be easily processed by the remaining modules. *Passage retrieval* module accomplishes 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

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 allow locating the documents that are likely to contain the answer.

These processes are performed by using a simple manually developed set of lexical patterns. Each pattern is associated with its corresponding expected

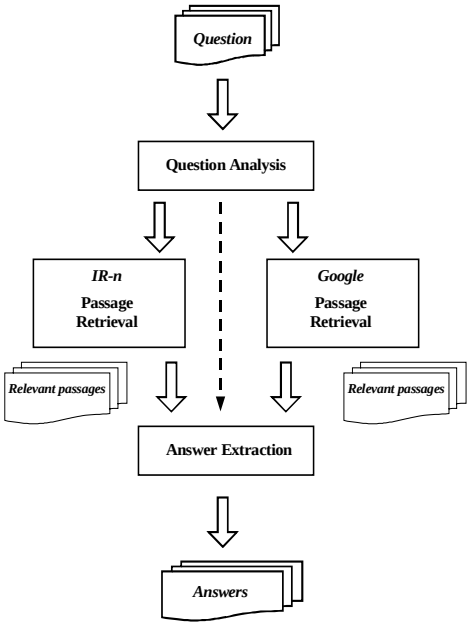

Fig. 1. System architecture

answer type. This way, 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 to the matched pattern. As our system lacks of 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 question analysis stage for test questions 002, 006 and 103.

2.2 Passage retrieval

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

IR-n system is a passage retrieval system that uses groups of contiguous sentences as unit of information. From 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, 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

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

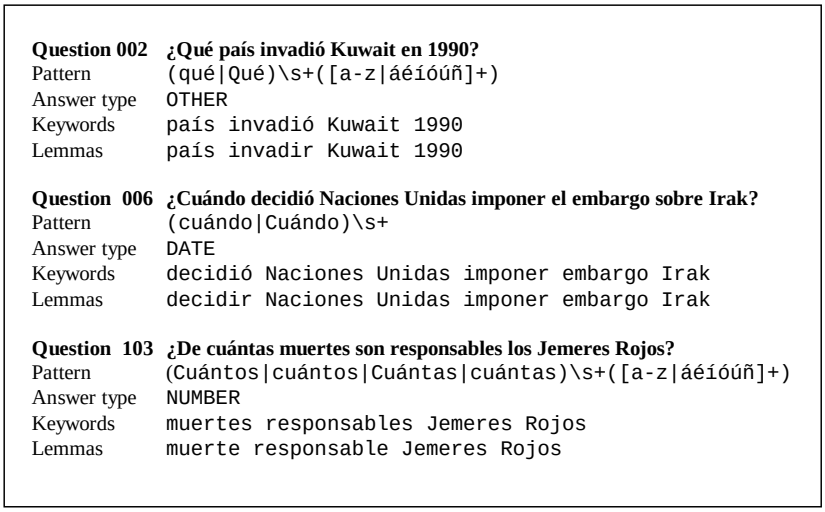

Fig. 2. Question analysis example

from the EFE document database. These passages are made up by text snippets of 2 sentences length. Second, the same keyword list (without being lemmatised) is posed to Google Internet search engine. Relevant documents are not downloaded. For efficiency considerations, the system only selects the 50 best short summaries returned in Google main retrieval pages. Figure 3 shows examples of retrieved passages for question 103. In this example question keywords found in relevant passages are underlined.

2.3 Answer extraction

This module processes both sets of passages selected at passage retrieval stage (IR-n and Google) in order to detect and extract the three more 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 times 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*).
2. *Candidate answer selection.* Candidate answers are selected from relevant sentences.
   (a) Relevant sentences are tagged using MACO lemmatizer.

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

(b) Quantities, dates and proper noun sequences are detected and they are merged into unique 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 of 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 unique expression that is scored according to the number of times this candidate appears 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.
4. *Web evidence addition.* All previous processes may be optionally performed in parallel for retrieving answers from web documents. Therefore, at this moment the system has two lists of candidate answers: one obtained from EFE document set and another from available Spanish web documents. If

web retrieval has been activated, candidate answer lists are merged. This process consists on increasing answer frequency of EFE list candidates by adding their corresponding frequency values obtained on web list. This way, candidates appearing only in 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 they have been found in (*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. Final answer score is computed as follows:

$$answer\_score = sentence\_score \cdot answer\_frequency \quad (1)$$

Answers are then ranked accordingly to their answer score and 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 final answer.

3 Results

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

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

Result analysis may not be as conclusive as we would desire mainly due to the simplicity of our approach. Besides, the lack of the correct answers for test questions at this moment do not allow us to perform a correct error analysis. Anyway, results obtained show that using the World Wide Web as external resource increases the percentage of correct answers retrieved in five points. This fact confirms that QA systems performance for other languages 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 of future work to be investigated. Among them, we can select the following ones:

- 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 systems ability of relating questions with their respective answers characteristics. Consequently, we need to develop a broad answer taxonomy that enables multilingual answer type classification. Probably, using EuroWordNet[4] semantic net structure.
- Passage Retrieval. An enhanced question analysis will improve passage retrieval performance by including question expansion techniques that enable retrieving passages including relevant information expressed with terms that are different (but equivalent) to those used for question formulation.
- Answer Extraction. Integrating named-entity taggers. 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 allows to narrow down the number of candidates to be considered for answering a question.

Even though all these lines need to be investigated, it is important to remark that this investigation needs to be developed from a multilingual perspective. That is, future investigations need to address language-dependent and language-independent modules detection and combination with the main long-term objective of developing a whole system capable of performing multilingual question answering.

References

1. Jordi Atserias, Josep Carmona, Irene Castellón, Sergi Cervell, Montse Civit, Lluís Màrquez, M.A. Martí, Lluís Padró, Roser Placer, Horacio Rodríguez, Mariona Taulé, and Jordi 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. John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, and Ralph Weishedel. 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

3. Fernando Llopis, José L. Vicedo, and Antonio 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.
4. José Luis Vicedo and Antonio Ferrández. A semantic approach to Question Answering systems. In *Ninth Text REtrieval Conference*, volume 500-249 of *NIST Special Publication*, pages 511–516, Gaithersburg, USA, nov 2000. National Institute of Standards and Technology.
5. José Luis Vicedo, Antonio Ferrández, and Fernando Llopis. University of Alicante at TREC-10. In *Tenth Text REtrieval Conference*, volume 500-250 of *NIST Special Publication*, Gaithersburg, USA, nov 2001. National Institute of Standards and Technology.
6. José Luis Vicedo, Fernando Llopis, and Antonio Ferrández. University of Alicante Experiments at TREC-2002. In *Eleventh Text REtrieval Conference*, volume 500-251 of *NIST Special Publication*, Gaithersburg, USA, nov 2002. National Institute of Standards and Technology.