



Combining Wikipedia and Newswire Texts for Question Answering in Spanish*

César de Pablo-Sánchez¹, José L. Martínez-Fernández^{1,2},
Ana González-Ledesma³, Doaa Samy¹, Paloma Martínez¹,
Antonio Moreno-Sandoval³, and Harith Al-Jumaily¹

¹ Universidad Carlos III de Madrid

{cdepablo, dsamy, pmf, haljumai}@inf.uc3m.es

² DAEDALUS, Data, Decisions and Systems S.A.

jmartinez@daedalus.es

³ Universidad Autónoma de Madrid

{ana, sandoval}@maria.111f.uam.es

Abstract. This paper describes the adaptations of the MIRACLE group QA system in order to participate in the Spanish monolingual question answering task at QA@CLEF 2007. A system, initially developed for the EFE collection, was reused for Wikipedia. Answers from both collections were combined using temporal information extracted from questions and collections. Reusing the EFE subsystem has proven not feasible, and questions with answers only in Wikipedia have obtained low accuracy. Besides, a co-reference module based on heuristics was introduced for processing topic-related questions. This module achieves good coverage in different situations but it is hindered by the moderate accuracy of the base system and the chaining of incorrect answers.

1 Introduction

MIRACLE team submitted a run for the Spanish monolingual QA subtask at CLEF [4] that included as innovations: Wikipedia as an additional collection and the move towards topic related questions. Our basic QA system uses a pipeline architecture [3] and is based on Information Extraction. Most successful systems for Spanish have opted either for a similar strategy like Priberam [1] or text-mining like INAOE [5]. Our aim was to test the adaption of the QA system to other collections. Therefore we reused the basic QA system and developed a new module for merging answers based on temporal information. Finally, to cope with topic-related questions we developed linguistically motivated heuristics to identify the focus of a question and test their accuracy.

The rest of the paper is structured as follows, the next section describes the system architecture focusing on the new modules. Section 3 introduces the results

* This work has been partially supported by the Regional Government of Madrid under the Research Network MAVIR (S-0505/TIC-0267) and projects by the Spanish Ministry of Education and Science (TIN2004/07083, TIN2004-07588-C03-02, TIN2007-67407-C03-01).

and a preliminary analysis of errors. Conclusions and directions for future work are discussed in Section 4.

2 System Overview

The architecture of the system (Figure 1) is similar to the multi-source architecture proposed in [2]. It is composed of two streams, one for each of the collections; EFE or Wikipedia. Each stream produces a ranked list of answers that are merged and combined by the Answer Source Mixer, a new module included for this evaluation. Question Analysis includes a new shared module for managing topic identification, context and anaphora resolution in topic-related question series. The basic system has been described in earlier participations [3] and it performs two kinds of operations; offline operations like indexing and document analysis and online operations like question analysis, sentence retrieval and answer selection.

2.1 Topic Identification in Topic-Related Questions

Introducing topic related questions requires a method to solve referential expressions that appear between questions and answers in the same question group. The system processes the first question and generates a set of candidates including the topic, the focus and the expected answer. A few rules that cover the most common cases are implemented to select the best topic for the question group. Rules use information available through question analysis and simplified assumptions about the syntactic structure of the questions.

The rules to locate the topic for a question group are :

- Answers of NUMEX subtype (numbers and quantities) are ignored as topics for questions series. The topic of the question, usually the syntactic subject will be the topic of following questions.
- Questions asking for a definition like *¿Quién es George Bush?* will add the topic and the answer (*presidente de los Estados Unidos*) to the group topic. An analog case occurs when we have questions like *¿Quién es el presidente de los Estados Unidos?*.
- Questions following the pattern *¿Qué NP * ?* ” like *¿Qué organización se fundó en 1995?*. In these cases the noun group following the interrogative article is the focus of the question. Both the answer and the focus are added to the group topic.
- For the rest of the cases we use the answer as the topic.

Once the topic for the group is identified, the rest of the questions use it as an additional relevant term in order to locate documents and filter relevant sentences. Obviously, there is a problem when the system is not able to find the right answer and this is the topic for the rest of the group.

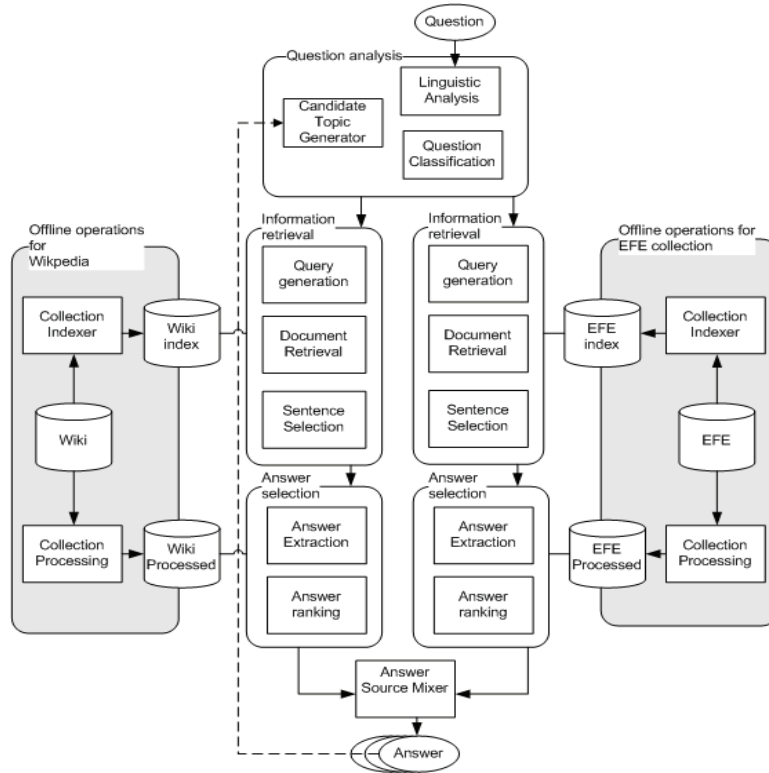


Fig. 1. MIRACLE 2007 System architecture

2.2 Combining EFE and Wikipedia Answers

The role of the Answer Source Mixer consists on combining answer candidates from the two different collections. We opted for a kind of 'semantic' reranking that takes into account the time period of collections, the verb tense and the time restrictions of the questions. In this way, no answer is really dropped from the candidates list but the list is reordered according to these clues. The heuristics used are the following:

- If the verb of the question appears in present tense preference is given to answers appearing in the Wikipedia collection.
- If the verb is in past tense and the question makes reference to the period covered by the EFE news collection, i.e., 1994 or 1995, then preference is given to answers from this collection.

3 Results and Error Analysis

Using the system described above, we have submitted one monolingual run for the Spanish subtask and obtained an overall accuracy of 15.00% (18.35% considering only factoids). Despite the inclusion of new sources of information like Wikipedia and the improvements carried in all preexisting modules, results obtained are lower than previous years [3]. However, this has been a general trend for all participants [4] due to the increasing difficulty of the task.

The complexity of the QA@CLEF task and the increasing number of stages that a question goes through in order to be answered, makes arduous any kind of error analysis. We reuse results presented in [4] and analyze them in the context of our system together with our own error analysis.

Regarding the main innovations, we have analyzed how the adaptations were able or not to solve them. There were 20 topic-related groups of questions with a total of 50 questions. The accuracy for the 170 questions that did not have to solve coreference was 15,29% while for the rest of the series is 13,33%. The difference is small and all additional errors except four are due to an incorrect selection of the first answer in the group.

In contrast, the analysis of the results for the different collections, EFE and Wikipedia, reveals that the source of the main decrease in accuracy is the strategy adopted for the latter one. For the 71 questions with answer in both collections the accuracy is 28,17%, slightly better than previous evaluations. When the answer could only be found in Wikipedia (114 questions), the accuracy decreases to 7.89%. This is specially acute for definitional questions whose accuracy dropped to 3,13%. This reveals that the system strategies have been overadapted to the EFE collection over the years, for example with heuristics like pronominals for definitions.

4 Conclusions and Future Work

Result analysis shows that the source of most problems appear in the Wikipedia stream where we applied the same strategies used in EFE with little success. The module for coreference resolution is effective even if it uses few heuristics. In contrast, the greater contribution of errors is due to the low accuracy at finding the first answer. Alongside the improvement in general performance we plan to study methods to cope with several candidate answers and uncertainty when answering series of topic-related questions.

References

1. Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., Vidal, D.: Priberam's question answering system in a cross-language environment. In: *Evaluation of Multilingual and Multi-modal Information Retrieval*, pp. 300–309 (2007)
2. Chu-Carroll, J., Prager, J.M., Welty, C.A., Czuba, K., Ferrucci, D.A.: A Multi-Strategy and Multi-Source Approach to Question Answering. In: *TREC (2002)*
3. de Pablo-Sánchez, C., González-Ledesma, A., Moreno-Sandoval, A., Vicente-Díez, M.: MIRACLE experiments in QA@CLEF 2006 in Spanish: main task, real-time QA and exploratory QA using Wikipedia (WiQA) (2007)
4. Giampiccolo, D., et al.: Overview of the CLEF 2007 Multilingual Question Answering Track (2008)
5. Juárez-González, A., Téllez-Valero, A., Denicia-Carral, C., Montes-y-Gómez, M.: Using machine learning and text mining in question answering. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006*. LNCS, vol. 4730, pp. 415–423. Springer, Heidelberg (2007)