MIRACLE Question Answering System for Spanish at CLEF 2007*

César de Pablo-Sánchez, José Luis Martínez,
Ana García-Ledesma, Doaa Samy, Paloma Martínez,
Antonio Moreno-Sandoval, Harith Al-Jumaily
Universidad Carlos III de Madrid
{cdepablo,dsamy,pmf,haljumai}@inf.uc3m.es
Universidad Autónoma de Madrid
{ana,sandoval}@maria.lllf.uam.es
DAEDALUS, Data, Decisions and Systems S.A.
jmartinez@daedalus.es

Abstract

This paper describes the system developed by MIRACLE group to participate in the Spanish monolingual question answering task at QA@CLEF 2007. A basic subsystem, similar to our last year participation, was used separately for EFE and Wikipedia collection. Answers from the two subsystems are combined using temporal information from the questions and the collections. The system is also enhanced with a correference module that processes question series based on a few simple heuristics that constraint the structure of the dialogue. The analysis of the results show that the reuse of strategies for factoids is feasible but definitions would benefit from adaptation. Regarding questions series, our heuristics have good coverage but we should find alternatives to avoid error chaining from previous questions.

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

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids, Spanish

1 Introduction

QA@CLEF 2007 has introduced two innovations over last year evaluation. The first change consists on topic-related questions, a series of ordered questions about a common topic that simulates

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

the dialogue between a user and the system to obtain information related to the topic. In this dialogue the user could introduce anaphoric expressions to refer to mentioned entities or events that appear in previous answers or questions. The second innovation is related to the inclusion of the November 2006 dump of the Wikipedia [14] as a source of answers in addition to the classic newspaper collections.

MIRACLE submitted a run for the Spanish monolingual task using a system that was enhanced to answer questions from Wikipedia and EFE collections. Each subsystem results were combined in a unified ranked list. The system also included a module that handles topic related questions. It identifies the topic and use it to enhance the representation of the following questions. The basic QA system was based on the architecture of our last year submission [6] although almost all components have evolved since then. This system was based on the use of filters for semantic information and therefore is tailored for factual questions. Last year we also tried to improve it for temporally restricted questions. An additional requirement has been to develop a fast system that could be competitive in real time with a classic information retrieval system.

This paper is structured as follow, the next section decribes the system architecture with special attention to the new modules. Section 3 introduces the results and a preliminary analysis of the kind of errors that the system made. Conclusions and directions of future work to solve the main problems follow in Section 4.

2 System Overview

The architecture of the system used this year is presented in figure 1 and it is composed of two streams for each of the sources similar to [4, 3]. The first stream uses the EFE newswire collection as a source of answers while the second uses Wikipedia. Each stream produces a ranked list of answers that are merged and combined by the Answer Source Mixer component described below. The two QA streams share a similar basic pipeline architecture and work as an independent QA system, with different configuration parameters, collections, etc. The way we perform question analysis is common to the two streams and therefore, when the two streams are composed, this module is shared. Another new common module complements question analysis for managing context and anaphora resolution in topic-related question series.

2.1 Basic system architecture

The basic system follows the classic pipeline architecture used by many other QA systems [8, 11]. The different operations are split between those that are performed online and offline.

2.1.1 Offline operations

These operations are performed during the preparation of the collection to speed up online document retrieval and answer extraction.

- **Collection indexing**. Collections are indexed at the word level using Lucene [1]. They are processed to extract the text that will be indexed. In the case of Wikipedia we remove format and links and therefore we do not use this information for relevance. Documents are indexed after tokenization, stopword removal and stemming based on Snowball [10].
- **Collection processing**. In order to speed the process of answer extraction we can enable the use of preprocessed collections. In this case, collections are analyzed using the output of language tools or services. For Spanish we use DAEDALUS STILUS[5] analyzers that provide tokenization, sentence detection and token analysis. Token analysis include detailed part of speech analysis, lemmatization and semantic information. STILUS has been improved from last year to support part of speech tagging. Regarding the semantic information, we have use STILUS Named Entities tagging, that is based on linguistic resources organized in a classification inspired by Sekine's typology[13]. The processed collection is stored on disk. Without compression the collection is about 10 times larger than the corresponding original.

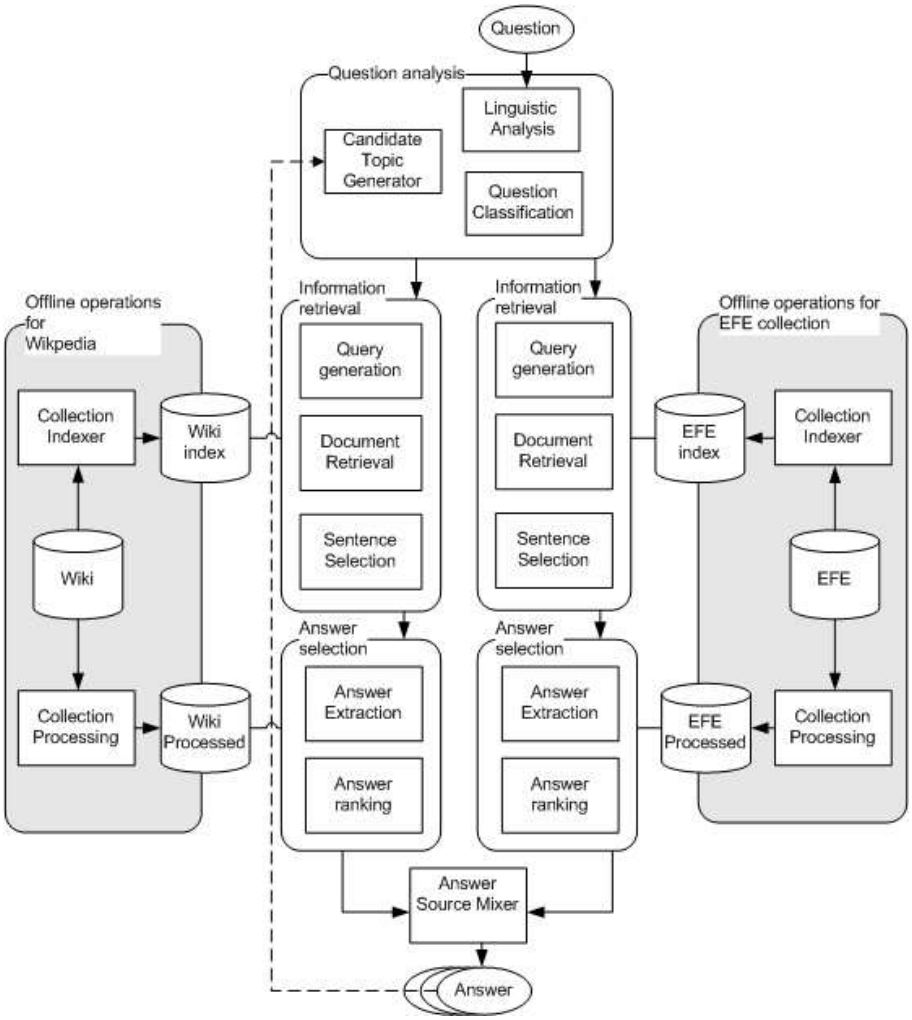

Figure 1: MIRACLE 2007 System architecture

2.1.2 Online operations

Online operations has been described in detail in previous participations [6]. We outline the description of the three main modules here and remark only the main changes.

- **Question analysis**. Questions are analyzed using STILUS online and the module produce a custom representation of the question that includes focus, theme, relevant and query terms, question type and expected answer type (EAT).
- **Passage retrieval**. We have change this module to use Lucene as information retrieval engine. The question reresentation is used to build a query using Lucene syntax and relevant documents are retrieved. Lucene uses a vector model for document representation and cosine similarity for ranking. Documents are analyzed and only sentences that contain a number of relevant terms are selected for the next step.
- **Answer selection**. This module uses the question class and the expected answer type to select an appropiate filter to locate candidate answers. Extracted candidates are grouped if they are similar and ranked using a score that combines redundancy, sentence score and document score.

2.2 New modules

2.2.1 Group topic identification in question series

With the inclusion of topic related questions, the system needed a method to solve referential expressions that appear between questions and answers in the same series. The system processes the first question and generates a set of candidates with the topic, the focus and the future answer. We have implemented a few rules that we believed covered the most common cases to select the best topic for the whole group. The rules use the information available after question analysis and simplified assumptions about the syntactic structure of the questions. The rules to locate the topic for the question series are only applied to the first question:

- Answers of subtypes `NUMEX` (numbers and quantities) are ignored as topics for questions series. We believe it is improbable for a number, if not representing any other type of entity, to be a natural topic. We use the subject, the topic of this question, as the topic of the question series.
- Answers of subtypes `TIMEX` (dates, years, etc...) are also ignored as topics, but in this case it is because we believed that they fall outside the guidelines for this year. We use the same rule than for `NUMEX` so far. Using a temporal expresion as a topic is very natural. In fact, most temporally restricted questions and specially those that have an event as restrictions are naturally split into two questions. This strategy has already been used by [12] to answer this kind of questions. We would need to recognize referential expressions like *ese año (that year)*, *durante ese período (during that period)*, etc. and solve the reference correctly in a second step.
- The question ask for a definition like *Quién es George Bush? (Who is George Bush?)* will add the topic (Named Entity) and the answer *(presidente de los Estados Unidos)* to the group topic. We have a similar case when we have questions like *Quién es el presidente de los Estados Unidos? (Who is the president of the United States?)*.
- The question follows the pattern `Qué NP * ?  (What NP * ?)` " like *Qué organización se fundó en 1995? (Which organization was created in 1995?)*. In this cases the noun group that follow the interrogative article is the focus of the question. Both the answer and the focus would be added to the group topic. We should remark that this case is different from a question beginning with a preposition like *En qué lugar... ? (In which place...?)*.
- For the rest of the classes we use the answer as the topic for the rest of the group.

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. This is obviously a problem when the topic was the answer but the system was not able to find the right one.

These rules are based on the structure of the information seeking dialogue and how we introduce new topics in conversation. Our rules select the most promising candidate using the firt question and ignoring the rest. Nevertheless, it is posible to shift the topic by introducing a longer referential expresion like the examples mentioned for `TIMEX`. We plan to investigate how to modify our procedure to work in two steps, generating a ranking list of cantidates and selecting the best candidate depending on constraints expressed in the referential expression.

2.2.2 Combining EFE and Wikipedia answers

As already mentioned, this year there are two possible collections to find the right answer to a question, one is Wikipedia and the other is the EFE newswire collection for years 1994 and 1995. This means that there should be some automatic method to decide which of these sources should be more relevant to extract candidate answers to a given question. This is the role of the Source Mixer component, based on very simple heuristics:

- 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 years appearing in the question, then preference is given to answers from the EFE news collection.

The preference given to each answer is measured by two parameters, one referring to the verb tense factor and the other referring to the time period factor. In this way, no answer is really dropped from the candidates list but the list is reordered according to this clues. At the output of the Source Mixer component there is a list of candidates ordered according to their source and the information present in the question.

3 Results

3.1 Run description

Using the system described above we have submitted one monolingual run for the Spanish subtask of this edition. Evaluation results and combined measures are outlined in tables 1 and 2.

Table 1: Judged answers for submitted run

Name	**Right**	**Wrong**	**Inexact**	**Unsupp.**
MIRA071ESES	30	158	4	8

Table 2: Evaluation measures for submitted run

Name	**Acc@1(F)**	**Acc@1(T)**	**Acc@1(D)**	**Acc@1(All)**
MIRA071ESES	18.35	13.95	3.13	15.00

Results are disappointing in general, as they are lower than previous years results for almost all types, despite the inclusion of new sources like Wikipedia. Even though almost all modules have been improved, the overall accuracy is not improving. We believe that whether the question set is more difficult or the system was not tuned as much as needed. We do not show results for the ten list questions because we did not implement an specific strategy. The case of definitions is analyzed with more detail below. We believe that question series introduce additional complexity in the task and this is reflected in the results. If we ignore question series we obtain an accuracy around 18% for the rest of the questions, which supports this idea but reflect that even base results are not good.

3.2 Analysis of errors

We are analysing our results in order to estimate which parts of the system need further improvement. For the moment, we are only able to present a preliminary analysis of the contents of our submission that uses one answer per question with their supporting sentence and document. The document returned is correct in 44% of the results, which is a lower bound to measure the performance of the document retrieval system. It also includes errors caused by an incorrect selection of a document collection. Given a correct document, 31% of the sentences selected with the first answer do not really contain a correct answer. Even if the sentence is correct, the error selecting a correct string is about 47%. These kind of error accumulates incorrect selection of a candidate, incorrect extraction and incorrect identification of the expected answer type. Finally, at least 8

questions have found an unsupported answer, which signals that in those cases where the same answer string has been extracted from several sentences, the score to select the most representative answer is not working as well as expected. A more detailed analysis isolating the contribution of the main modules will be presented in the final version of this article.

Analyzing the behaviour of the system for different types we have detected that the accuracy of definition questions has dropped dramatically. We believe that the heuristics that we have defined to extract definitions in EFE do not work for Wikipedia. The system used to signal appositions, nominal phrases before Named Entities and expansion of acronyms as valid definitions of persons and organizations. In contrast, definitional sentences in Wikipedia usually are copulative, they have a longer distance between the defined object and a valid definition and they usually are placed in the beginning of the document. Unfortunately, we did not implement any special strategy for these questions. If we consider the rest of the types, the behaviour is similar to previous years, most of the factual questions with well defined Named Entities achieve a reasonable accuracy. Questions with EAT OTHER are in contrast much harder.

There were 20 topic related groups of questions that made a total of 50 questions. The system correctly answered 5 questions of this kind, from three different groups. This behaviour was expected as errors usually chain, specially in the case when the answer to the first question is not correct and this happens to be the topic. We have analyze the main source of errors in order to evaluate the correference component. Rules for topic detection are the source of errors only for three of the cases, and therefore seems to work reasonably well although there is room for some simple improvements. In one of these cases, the error is due to a question not correctly identified as a NUMEX type. The rest of the errors are due to incorrect identification of the first answer and the chaining of errors.

4 Conclusions and Future Work

We have presented the architecture of the MIRACLE QA system and the main modifications introduced to cope with new challenges in the CLEF@QA task. Using this system we have submitted one run for the monolingual Spanish subtask where results were lower than expected.

The analysis of the performance across types have signalled that for some type of questions the style of the text is an important issue. This is specially accute in the case of definitions. We have employed the same subsystem and strategies for the EFE and the Wikipedia collections with dissapointing results. The analysis of the errors have shown that the methods for document retrieval and candidate extraction could be adapted to improve the accuracy. We plan to use type speficic approaches like the ones used by [7, 9] and collection specific approaches to improve results for these types.

We have found that the module for correference resolution is effective even if it uses a limited amount of knowledge. In contrast, the greater contribution of errors is due to the low accuracy at the first answer. This is even more accute as a great deal of the questions that set the initial topic are definitional. Besides type specific approaches, we need to improve the correference method to consider more than one candidate answer and cope with uncertainty.

Another question that we have not evaluated thoroughly yet is the way we combine results. We use a semantic kind of combination that exploits the different time spans of the two collections. It is still unclear if these method is appropriate or whether techniques adapted from information retrieval from heterogeneus collections as in [2] would work better.

References

[1] Lucene webpage. http://lucene.apache.org/, August 2007.

[2] Rita M. Aceves-Perez, Manuel Montes y Gomez, and Luis Villasenor-Pineda. Fusion de respuestas en la busqueda de respuestas multilingue. *SEPLN, Sociedad Espaola para el Procesamiento del Lenguaje Natural*, 38:35–41, 2007.

[3] David Ahn, Valentin Jijkoun, Karin Mller, Maarten de Rijke, Stefan Schlobach, and Gilad Mishne. *Making Stone Soup: Evaluating a Recall-Oriented Multi-stream Question Answering System for Dutch*. 2005.

[4] Jennifer Chu-Carroll, John M. Prager, Christopher A. Welty, Krzysztof Czuba, and David A. Ferrucci. A multi-strategy and multi-source approach to question answering. In *TREC*, 2002.

[5] DAEDALUS. Stilus website. On line http://www.daedalus.es, July 2007.

[6] Cesar de Pablo-Sanchez, Ana Gonzalez-Ledesma, Antonio Moreno-Sandoval, and Maria-Teresa Vicente-Diez. MIRACLE experiments in QA@CLEF 2006 in spanish: main task, real-time QA and exploratory QA using wikipedia (WiQA). 2007.

[7] Abdessamad Echihabi, Eduard Hovy, and Michael Fleischman. Offline strategies for online question answering:. 2003.

[8] Dan I. Moldovan, Sanda M. Harabagiu, Marius Pasca, Rada Mihalcea, Roxana Girju, Richard Goodrum, and Vasile Rus. The structure and performance of an open-domain question answering system. In *ACL*, 2000.

[9] Manuel Montes-y Gomez, Luis Villaseñor Pineda, Manuel Pérez-Coutiño, José Manuel Gómez-Soriano, Emilio Sanchís-Arnal, and Paolo Rosso. A full data-driven system for multiple language question answering. *: Accessing Multilingual Information Repositories*, pages 420–428, 2006.

[10] Martin Porter. Snowball stemmers and resources website. http://www.snowball.tartarus.org, July 2007.

[11] John Prager, Eric Brown, Anni Coden, and Dragomir Radev. Question-answering by predictive annotation. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Question Answering, pages 184–191, 2000.

[12] E. Saquete, J.L. Vicedo, P. Martínez-Barco, R. Muñoz, and F. Llopis. Evaluation of complex temporal questions in clef-qa. *: Multilingual Information Access for Text, Speech and Images*, pages 591–596, 2005.

[13] Satoshi Sekine. Sekine's extended named entity hierarchy. On line http://nlp.cs.nyu.edu/ene/, August 2007.

[14] November 2006 dump of wikipedia. http://download.wikimedia.org/images/archive/eswiki/20061202/pages-articles.xml.bz2, July 2007.