# Are Passages Enough?
# The MIRACLE Team Participation at QA@CLEF2009

María Teresa Vicente-Díez, César de Pablo-Sánchez, Paloma Martínez,
Julián Moreno Schneider, Marta Garrote Salazar
Universidad Carlos III de Madrid

{tvicente, cdepablo, pmf, jmschnei, mgarrote}@inf.uc3m.es

**Abstract**

This paper summarizes the participation of the MIRACLE team in the Multilingual Question Answering Track at CLEF 2009. In this campaign, we took part in the monolingual Spanish task at ResPubliQA@CLEF 2009 and submitted two runs. We have adapted our QA system which has been evaluated in EFE and Wikipedia to the new JRC-Acquis collection and the legal domain. We tested the use of answer filtering and ranking techniques to a base system using passage retrieval with no success. Our run using question analysis and passage retrieval obtained a global accuracy of 0.33 while the addition of an answer filtering step obtained 0.29. We provide an initial analysis of the results across the different questions types while we research the reason why it is difficult to leverage previous QA techniques. A different focus of our work has been on temporal reasoning applied to question answering and also detailed discussion of this issue in the new collection and analysis of the questions is provided.

## 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

## Keywords

Question Answering, Spanish, legal domain, temporal indexing, temporal normalization

## 1 Introduction

We describe the MIRACLE team participation in the ResPubliQA exercise at the Multilingual Question Answering Track at CLEF 2009. The MIRACLE team is a consortium formed by three universities from Madrid, (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) and DAEDALUS, a small and medium size enterprise (SME). We submitted two runs for the Spanish monolingual subtask which summarize our attempts to adapt our QA system to the new requirements of the task.

This year, the main task departed from previous exercises in an attempt to explore new domains, question types and multilingual experiments. The change in application domain has been triggered by the use of the JRC-Acquis document collection which is formed by European legislation translated in several EU languages. This fact raises the problem of dealing with legal language which includes richer terminology and is considerably more complex than news or academic language used in EFE and Wikipedia collections. Moreover, new kind of information needs are required to be solved which has motivated the inclusion of question asking for objectives, motivations, procedures, etc. in addition to the traditional factual and definitional questions. The new types of questions often required longer answers and therefore the expected response of the system has been fixed again at the paragraph level. Nevertheless it should be possible to take advantage of answer selection techniques developed in previous campaigns. This has been in fact one of the hypothesis we would like to test with our participation. Unfortunately, our experiments in this line have not been successful and we have not found configurations that performed substantially better than our baseline. A different aspect of our work has centered on the use of temporal information in the process of QA and we report results for different indexing configurations. Finally, a global objective was to enlarge the capabilities of the QA system and advance towards an architecture that allows domain adaptation and multilingual processing.

The rest of the paper is structured as follows, the second section describes the system architecture with special attention paid to the novelties introduced this year, Section 3 introduces the submitted runs and analyzes the results. Finally, conclusions and future work are presented in Section 4.

## 2 System Description

The system architecture is similar to our previous system [2] and is based on a pipeline which analyzes questions, retrieves documents and performs answer extraction based on linguistic and semantic information. Different strategies can be used depending on the type of the question and the expected type of the answer. The architectural schema is shown in Figure 1. A number of modules have been modified, extended or reorganized in order to adjust for the requirements of the task and the legal domain. Other modules have been included to carry new experiments.

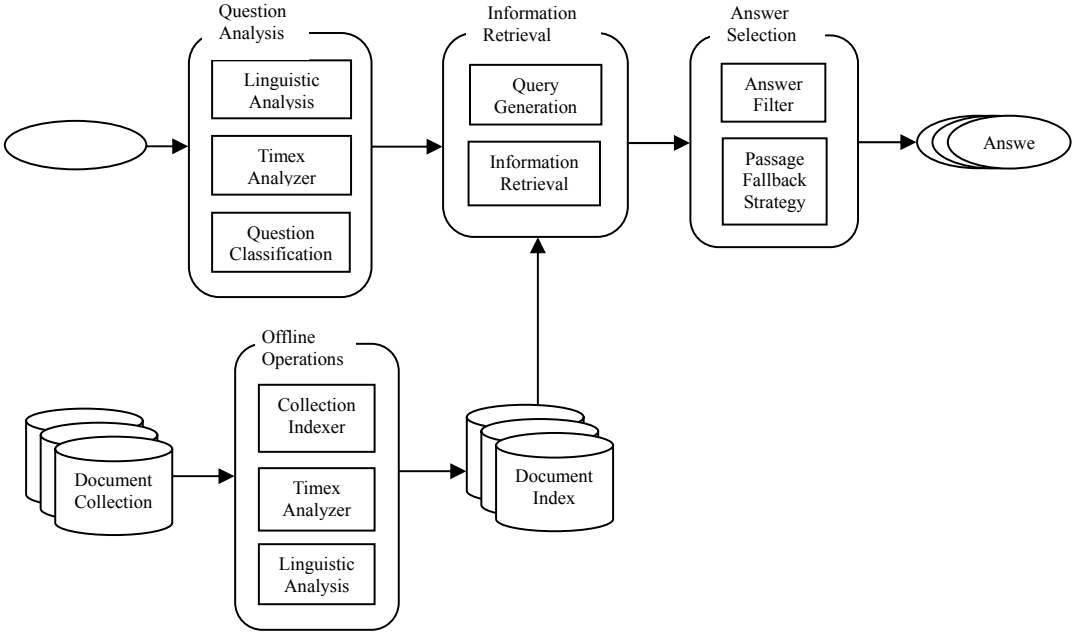


*Figure 1: MIRACLE 2009 system architecture*

The main changes performed in the system are outlined:

- Adding parsers for the new collections as well as supporting the indexing of passages.
- The evaluation procedure was modified to work for passages and a fallback strategy for passages was included.
- New rules have been developed for Question Analysis, Question Classification and Answer Filtering for the legal domain using the development set.
- Query generation has been adapted to the domain and page heuristics for Wikipedia removed.
- Temporal Management was added to normalize temporal expressions and integrated into language analysis and indexing routines.
- New functionality for mining acronyms offline and add them to query generation.
- The ranking module was redesigned for modularity.

### *Indexes*

Indexes are really important for QA as obtaining a good retrieval subsystem can considerably improve the final results of the QA system. Due to the change in the document collection, all IR indexes have been newly created using Lucene as the retrieval engine. To accomplish the task of storing the relevant information as appropriately as needed, we have designed two different document types or indexing units:

- *Document*, where all the information related to title, note and the text of a collection file is stored.
- *Paragraph*, which store each paragraph, the title and the notes in a different document. Lucene uses a length document normalization term in the retrieval score which was arguably of no help in the case of

paragraph scoring because paragraphs are expected to have more uniform lengths. Both types of indexes, with length normalization and without were tested.

In all our experiments previous to the submission the paragraph or passage index worked better than the document index. Besides, we also created different index types regarding the analysis, characterized by the linguistic analyzer used in each case:

- *Simple Index*, where the text analyzer used is a simple analyzer adapted for Spanish. It makes grammar based parsing, stems words using a snowball-generated stemmer, removes stop words, replaces accented characters in the ISO Latin 1 character set and converts text into lower case. All the texts are stored in the same field: *text*.
- *Temporal Index*, which adds a recognition and normalization of time expressions. These time expressions are normalized and included in the index. Texts are also stored also in the field *text*.

Finally, other modifications required the query generation process to be changed to use the same analyzer that was used to create the index.

The idea of a rule engine, was initially considered for classifying question types; later, it has also been used not only in the Question Classification module, but also in the Answer Filter, Timex Analyzer and Topic Detection ones [2]. The rules have a left part that expresses a pattern and a right part specifying the actions to be taken each time the pattern is found. The pattern could refer to lexical, syntactic and/or semantic elements.

The change of linguistic domain meant some changes in the new rules. Below, we present an example of a new rule, developed to handle the extraction of definitions on this year corpus:

*Figure 2: Example of rule for answer extraction*

```
RULE("definition")
  EXISTENTIAL QUESTION TYPE ("DEFINITION") AND
  WORD_I(N, OBTAIN_FOCUS()) AND
  (WORD_I(N+1, ":") OR
   WORD_I(N+1, "\"") AND
   WORD_I(N-1, "\"") OR
   WORD_I(N+1, "\"")  AND
   WORD_I(N+2, ":") AND
   WOD_I(N-1, "\""))
THEN
   ANSWER_EXTRACTION(0,POS_LAST_TOKEN());
END
```

This rule has been created to detect the topic in definition questions. In most of them, the topic in the answer paragraph was written in quotation marks and/or followed by colon. This rule locates the topic of the question and looks for it in the source documents.

### *Temporal Management*

Some authors have defined the temporal question answering (TQA) as the specialization of the QA task in which questions have some features that denote temporality [4], as well as a means for providing short and focused answers to temporal information needs formulated in natural language [6]. Previous work has already faced up to this problem for the treatment of other languages, like in [7] or [8], or also in Spanish [3]. Temporal questions can be classified into 2 main categories according to the role of temporality in their resolution:

- Temporally Restricted (TR) questions are those containing some time restriction: *"¿Qué resolución fue adoptada por el Consejo el 10 de octubre de 1994?" ("What resolution was adopted by the Council on 10 October 1994?*")
- Questions with a Timex Answer (TA) are those whose target is a temporal expression or a date: *"¿Cuándo empieza la campaña anual de comercio de cereales?" ("When does the marketing year for cereals begin?"*)

In this campaign, temporal management preserves the approach taken by the MIRACLE QA system participating in CLEF 2008 [2]. This decision is based on later complementary work that was made in order to evaluate the QA system performance versus a baseline system without temporal management capabilities [9]. The experiments showed that additional temporal information management can quantitative and qualitatively benefit the results. This led us to predict that the use of such strategies could enrich future developments.

Several adjustments were made in the temporal expressions recognition, resolution and normalization integrated system to enhance its coverage on the new collections. Similarly to the previous version, the date of creation of each document is adopted as the reference date, needed to resolve the relative expressions that contains. In JRC-Acquis documents this information is provided by the "*date.created*" attribute.
Question analysis, indexes generation and answer selection modules have been considered potentially more influential for achieving better results by means of the application of temporal management. They have been slightly adapted to the requirements of this year's competition, keeping the essence of their functionality.

- During question analysis process, queries, including those with temporal features, are classified, distinguishing between TR and TA queries. If a TA query is detected, it determines the granularity of the expected answer (complete date, only year, month, etc.).
- The answer selector is involved in two directions: in the case of TA queries, the module must favour a temporal answer, whereas if it manages TR queries, it applies extraction rules based on the temporal inference mechanism and demotes the candidates not fulfilling the temporal restrictions.

As a novelty, this year we have created more sophisticated indexes according to the paragraph retrieval approach of the competition. In some configurations, the normalized resolution of temporal expressions is included in the index instead of the expression itself. The main objective is to assess the behaviour of the QA system using different index configurations, mainly focusing on the temporal queries of the collection.

### *Acronym mining*

Due to the nature of the collection, a large number of questions were expected to be expansion of acronyms, especially about organizations. On the other hand, the recall of the information retrieval step could be improved by including the acronym and their expansion in the query.

We implemented a simple offline procedure to mine acronyms by scanning the collection and searching for a pattern which introduces a new entity and provides their acronym between parentheses. Then, results are filtered in order to increase their precision. First, only those associations that occur at least twice in the corpus are considered. As parentheses often convey other relations like persons and their country of origin, another filter removed countries (Spain) and their acronyms (ES) from the list. Finally, some few frequent mistakes were manually removed and acronyms with more than one expansion were also checked.

Once we have cleaned the file, we index the acronyms and their expansions separately to be able to search by acronym or by expansion.

The index is used in two different places in the QA system:

- Query Generation, where it analyzes the question and adds searching terms to the query that is sent to the document collection index.
- Answer Filter, where it analyzes the text extracted from the paragraph to determine if that paragraph contains the acronym (or the expansion) and if so, identifies the paragraph as correct answer.

### *Answer Filter and Passage Fallback Strategy*

This module, previously called Answer Extractor, process the result list from the information retrieval module and selected chunks to form a possible candidate answer. In previous years, this module was designed to extract answers selected from the document. In this campaign, the answer must be the complete text of a paragraph therefore, this year the module works as a filter which removes passages with no answers. The kind of linguistic rules used last year to perform answer extraction has been adapted and new rules to detect acronyms, definitions as expressed in the new corpora and new rules for temporal questions have been developed.

The possibility of getting no answer from the answer filter led to the development of a module that simply creates answers from the retrieved documents. This module is called Passage Fallback Strategy. It takes the documents returned by the information retrieval module and generates an answer from every document. The way

of generating the indexes (concretely the paragraph index) makes possible the functionality of this module.

### *Evaluation module*

Evaluation is a paramount part of the development process of the QA system. In order to develop and test the system the English development test provided by CLEF organizers was translated to Spanish and a small gold-standard with answers was developed. Mean Reciprocal Rank (MRR) and Confidence Weighted Score (CWS) were consistently used to compare the outputs of the different configurations with the development gold standard. Periodically, the output and the XML logs of different executions were manually inspected to complete the gold standard and to detect integration problems.

## 3 Experiments and results

We submitted two runs for the monolingual Spanish task. They correspond to the configurations of the system that yielded best results during our development using the translated question set. Paradoxically, both runs match with the simplest configurations that we have tested.

- Baseline (mira091eses): The system is based on passage retrieval using the simple index. Question analysis is performed to generate queries and the acronym expansion is used.
- Baseline + Answer Filter (mira092eses): Adds answer filtering and the passage fallback strategy after the previous passage retrieval.

A number of additional configurations were also tested but no improvements over the baseline were found consistently. In fact, most of the additions seem to produce worse results on our development test. We considered different functions for Answer Ranking and Passage Re-ranking which we have tested for previous participations and some new ones. Different passage length normalization strategies were also applied to the indexes. Finally, a great deal of effort was devoted to the treatment of temporal expressions in question analysis, indexing and extraction and more detailed experiments are presented below.

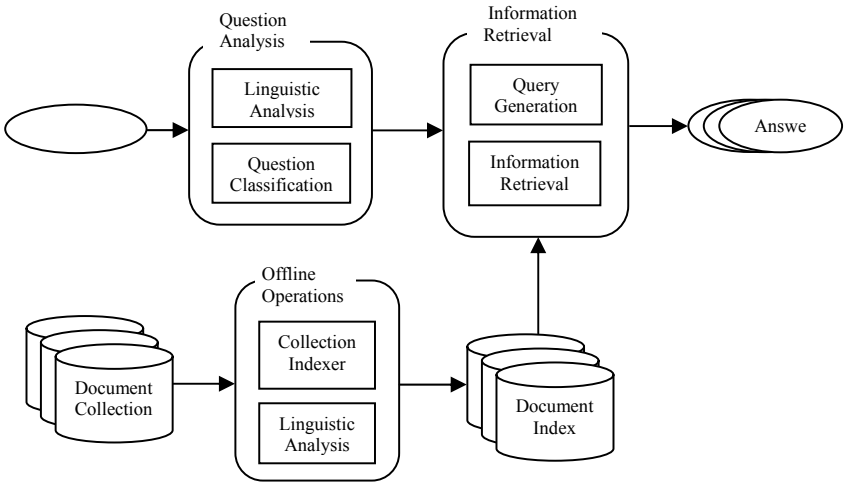


*Figure 3: mira091eses configuration (BL)*

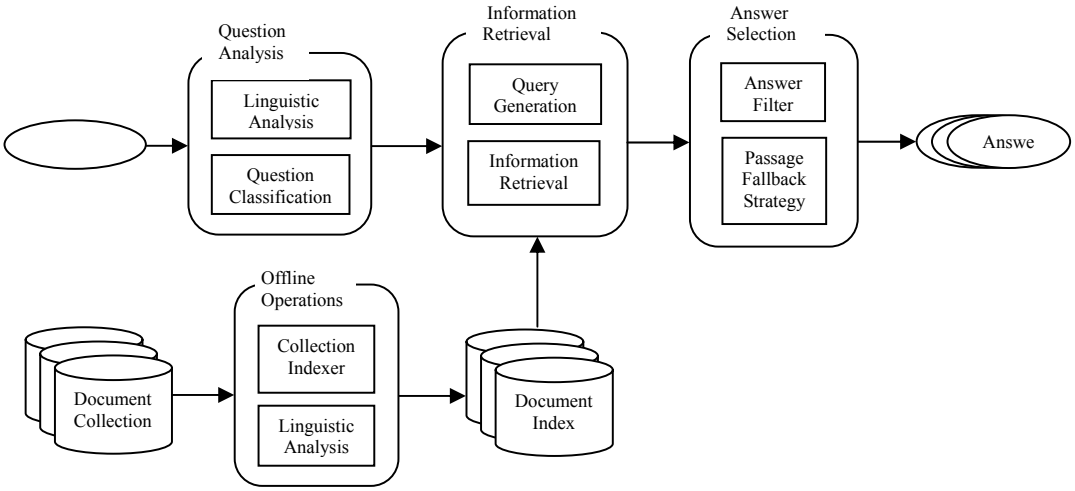


*Figure 4: mira092eses configuration, including answer filtering (BL+AF)*

Evaluation figures are detailed in Table 1. Answer accuracy has been calculated as the ratio of questions correctly answered to the total number of questions. Only the first candidate answer is considered, rejecting the rest of possibilities.

| Name | Right | Wrong | Unanswered with Right Candidate Answer | Unanswered with Wrong Candidate Answer | Unanswered with Empty Candidate | Overall accuracy | Proportion of answers correctly discarded | c@1 measure |
|---|---|---|---|---|---|---|---|---|
| mira091eses | 161 | 339 | 0 | 0 | 0 | **0.32** | 0 | 0.32 |
| mira092eses | 147 | 352 | 0 | 0 | 1 | **0.29** | 0 | 0.29 |

*Table 1: Results for submitted runs*

The results on the CLEF09 test set show similar conclusions to those we obtained during our development process, the baseline system using passage retrieval is hard to beat and in fact our second run provide lower accuracy. As in the case of our development experiments there are changes for individual answers of a number of questions but the overall effect is not positive.

After the evaluation, and using the larger test set of 500 questions we have decided to carry a class based analysis in order to understand the causes behind our unfruitful efforts. We have manually annotated the questions and grouped them in 6 main question types. In contrast with our expectations, the performance of the second submitted run is also worse for the factual and definition questions. As we have considered these questions types in previous evaluations we expected to have better coverage in the Answer Filter and therefore an improvement. Similar behaviour has been observed across answer types for factual questions, being the class of TIMEX questions the only where the more complex configuration really improves.

Our analysis of the errors show that further work is needed to be able to cope with the complexities of the domain. For example, questions are in general more complex and include a large number of domain specific terminologies that our question analysis rules do not handle correctly. The process of finding the focus of the question which is crucial for question classification is specially error prone. Answer Extraction needs also further adaptation to the domain for factual questions as the typology of NE and generalized NE has not wide coverage. Problems with definitions are rooted more deeply and probably require the use of different specialized retrieval strategies. This year evidence along with previous experiments seems to support that definitions depend deeply on the stylistics of the domain. Finally, new question types would require further study of techniques that help to improve the classification of passages as bearing procedures, objectives, etc.

| Question Type | mira091eses | mira092eses | TOTAL | mira091eses Accuracy | mira092eses Accuracy |
|---|---|---|---|---|---|
| | BL | BL-AF | | BL | BL-AF |
| FACTUAL | 54 | 48 | 123 | 0.44 | 0.39 |
| PROCEDURE | 22 | 15 | 76 | 0.28 | 0.20 |
| CAUSE | 43 | 44 | 102 | 0.42 | 0.43 |
| REQUIREMENT | 5 | 5 | 16 | 0.31 | 0.31 |
| DEFINITION | 16 | 12 | 106 | 0.16 | 0.11 |
| OBJECTIVE | 21 | 23 | 77 | 0.27 | 0.30 |
| ALL | 161 | 147 | 500 | 0.32 | 0.29 |
| ALL - FACTUAL | 107 | 99 | 377 | 0.28 | 0.26 |

*Table 2: An analysis of runs by question type*

### *Evaluation of temporal questions*

With the aim of evaluating the temporal management capabilities of the QA system, we decided to extract the temporal questions from the whole corpus. 46 out of 500 queries denote temporal information, that means a 9,20% over the total. 24 of them are TR questions, whereas TA queries are 22 (4,80% and 4,40% out of the total, respectively). This subset has been studied, evaluating the correctness of the returned answers by two different configurations of the QA system. The results are presented in Table 3.

| Name | Temporal Questions (TR + TA) | Temporally Restricted (TR) | Timex Answer (TA) |
|---|---|---|---|
| BL (mira091eses) | 0.43 | 0.42 | 0.45 |
| BL-AF (mira092eses) | 0.48 | 0.37 | 0.59 |
| DA-BL (run1 configuration) | 0,28 | 0,21 | 0,36 |
| DA-BL-AF (run2 configuration) | 0,37 | 0,21 | 0,54 |

*Table 3: Results for temporal questions in the submitted runs and other configurations*

As we can observe, better figures are obtained by the set of TQ in both runs. There is no significant difference between TA and TR queries in the first run, while in the second one they achieve a difference of 22%. In our opinion, the second configuration, with answer filtering and answer creation, enhances precision for TA queries, whereas for TR queries, temporal restrictions introduce noise that the system is not able to solve.

Non-submitted runs present similar configurations to the submitted ones, but adopting a different index generation and question analysis strategies. The approach consisted on the inclusion of normalized temporal expressions into the index, as well as in the question analysis process, aiming to increase recall. We tested the performance over the total corpus of questions, but worse results were achieved even if the study is restricted to temporal questions. Results are also presented in Table 3, which show no improvement regarding the submitted runs. Performance difference between TA and TR queries remains stable, since the system has a better response to questions without temporal restrictions. The lost of accuracy can be due to the lack of a more sophisticated inference mechanism at the time of retrieval, capable of reasoning with different granularities in normalized dates format [10]. In addition, we suspect that answer selection module is not filtering candidate answers properly, so current inference mechanism gives more weigh to paragraphs containing dates matching with restrictions in the query, while the rest of terms lose relevancy. Though relative dates present a low frequency in the collections, they are not being correctly solved, as reference date, taken from that of the documents creation, is always set to the same value.

## 4 Conclusion and Future Work

From our point of view, the new ResPubliQA exercise is a challenge for QA systems in two main facets of the problem domain adaptation and multilinguality. This year our efforts have focused on the first problem where we

have ported the system and the techniques developed for EFE and Wikipedia to the new legal collection JRC-Acquis. However, our experiments, which are exemplified with the submitted runs, show that a system mainly based on passage retrieval performs quite well. Baseline passage retrieval results provided by the organizers [11] also support these. We are carrying further experiments using the larger test set in order to find how answer selection could help for ResPubliQA questions as well as the differences between passage retrieval alternatives.

Regarding our focus on temporal reasoning applied to QA we would explore how question temporal constraints can be integrated at other steps in the process. We expect to compare the effectiveness of temporal reasoning as constraints for filtering answers and for the purpose of re-ranking.

Finally, further work in the general architecture of the QA is expected to help in at least three areas: separation of domain knowledge from general techniques, adding different languages to the system and effective evaluation.

## Acknowledgements

This work has been partially supported by the Regional Government of Madrid by means of the Research Network MAVIR (S-0505/TIC/000267) and by the Spanish Ministry of Education by means of the project BRAVO (TIN2007-67407-C3-01)

## References

[1] Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006). ***The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages***. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

[2] Martínez-González, A., de Pablo-Sánchez, C., Polo-Bayo, C., Vicente-Díez, M.T., Martinez-Fernández, P., Martínez-Fernández, J.L. 2008. ***The MIRACLE Team at the CLEF 2008 Multilingual Question Answering Track***. In Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Series LNCS (to appear)

[3] ***Apache Lucene project***. The Apache Software Foundation. http://lucene.apache.org/, visited 30/07/2009.

[4] Saquete, E. ***Resolución de Información Temporal y su Aplicación a la Búsqueda de Respuestas***. 2005. Thesis in Computer Science, Universidad de Alicante.

[5] Saquete, E., Martínez-Barco, P., Muñoz, R., Viñedo, JL. 2004. ***Splitting Complex Temporal Questions for Question Answering Systems***. In Proceedings of the ACL'2004 Conference, Barcelona, Spain.

[6] De Rijke et al. ***Inference for temporal question answering Project***. 2004-2007. OND1302977.

[7] Hartrumpf, S. and Leveling, J. 2006. ***University of Hagen at QA@CLEF 2006: Interpretation and normalization of temporal expressions***. In Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop. Alicante, Spain.

[8] Clark, C. and Moldovan, D. ***Temporally Relevant Answer Selection***. In Proceedings of the 2005 International Conference on Intelligence Analysis, May 2005.

[9] Vicente-Díez, M.T. y Martínez, P. ***Aplicación de técnicas de extracción de información temporal a los sistemas de búsqueda de respuestas***. Procesamiento del lenguaje natural. N. 42 (marzo 2009); pp.25-30.

[10] ISO8601:2004(E) ***Data elements and interchange formats – Information interchange – Representation of dates and times***. Third edition 2004

[11] Pérez j. , Garrido G. , Rodrigo A., Araujo L., Peñas A. ***Information Retrieval Baselines for the ResPubliQA task***. 2009. CLEF 2009 Working Notes.