# The TALP-QA System for Spanish at CLEF 2005

Daniel Ferrés, Samir Kanaan, Alicia Ageno, Edgar González,
Horacio Rodríguez, and Jordi Turmo

TALP Research Center, Software Department, Universitat Politècnica de Catalunya,
Jordi Girona 1-3, 08043 Barcelona, Spain
{dferres,skanaan}@lsi.upc.edu
http://www.lsi.upc.edu/~nlp

**Abstract.** This paper describes the TALP-QA system in the context of the CLEF 2005 Spanish Monolingual Question Answering (QA) evaluation task. TALP-QA is a multilingual open-domain QA system that processes both factoid (normal and temporally restricted) and definition questions. The approach to factoid questions is based on in-depth NLP tools and resources to create semantic information representation. Answers to definition questions are selected from the phrases that match a pattern from a manually constructed set of definitional patterns.

## 1 Introduction

This paper describes TALP-QA, a multilingual open-domain Question Answering (QA) system under development at UPC for the past 3 years. A first version of TALP-QA for Spanish was used to participate in the CLEF 2004 Spanish QA track (see [5]). From this version, a new version for English was built and was used in TREC 2004 [6]. The main changes of the system architecture with respect to the prototype used in the CLEF 2004 evaluation are: i) factual and definition questions are treated with different architectures, ii) new modules have been designed to deal with temporally restricted questions, iii) Named Entity Recognition and Classification (NERC) and the Question Classification modules have been improved.

In this paper the overall architecture of TALP-QA and its main components are briefly sketched, the reader can consult [5] and [6] for more in depth description of this architecture. Most of the paper describes with some details the improvements over the previous system that have been included for this evaluation. We also present an evaluation of the system used in the CLEF 2005 Spanish QA task for factoid, temporally restricted factoid, and definition questions.

## 2 Factual QA System

The system architecture for factual questions has three subsystems that are performed sequentially without feedback: Question Processing (QP), Passage Retrieval (PR) and Answer Extraction (AE). This section describes the three main subsystems and the Collection Pre-processing phase.

### 2.1 Collection Pre-processing

We pre-processed the document collection (EFE 1994 and EFE 1995) with linguistic tools (described in [5]) to mark the part-of-speech (POS) tags, lemmas, Named Entities (NE), and syntactic chunks. Then, we indexed the collection and we computed the *idf* weight at document level for the whole collection. We used the *Lucene*[1] Information Retrieval (IR) engine to create an index with two fields per document: i) the lemmatized text with NERC and syntactic information, ii) the original text (forms) with Named Entity Recognition and syntactic information.

### 2.2 Question Processing

A key point in QP is the Question Classification (QC) subtask. The results of QC in our previous attempt (in CLEF 2004) were rather low (only 58.33% accuracy). As was explained in [5] the low accuracy obtained is basically due to two facts: i) the dependence on errors of previous tasks [5], ii) the question classifier was trained with the manual translation of questions from TREC 8 and TREC 9 (about 900 questions). The classifier performs better in English (74% (171/230)) than in Spanish (58.33% (105/180)), probably due to the artificial origin of the training material.

We decided to build a new QP module with two objectives: i) improving the accuracy of our QC component and ii) providing better material for allowing a more accurate semantic pre-processing of the question. The QP module is structured into five components, we will describe next these components focusing in those having changed from our previous system (see [5] for details):

- **Question Pre-processing**. This subsystem is basically the same component of our previous system with some improvements. For CLEF 2005 (for Spanish) we used a set of general purpose tools produced by the UPC NLP group: *Freeling* [2], *ABIONET* [3], *Tacat* [1], *EuroWordNet* (EWN), and *Gazetteers* [5]. These tools are used for the linguistic processing of both questions and passages. The main improvements on these tools refer to:
  - **Geographical gazetteers**: Due to the limited amount of context in questions, the accuracy of our NER and NEC components suffers a severe fall, specially serious when dealing with locatives, (a 46% of NEC errors in the CLEF 2004 questions analysis were related with locatives). For this reason, we used geographical gazetteers to improve the accuracy of the NEC task. The gazetteers used were: a subset of 126,941 non-ambiguous places from the GEOnet Names Server (GNS)[2], the *GeoWorldMap*[3] gazetteer with approximately 40,594 entries (countries, regions and important cities), *Albayzin Gazetteer* (a gazetteer of 758 place names of Spain existing in the speech corpus Albayzin [4]).

[1] `http://jakarta.apache.org/lucene`
[2] **GNS**. `http://gnswww.nima.mil/geonames/GNS/index.jsp`
[3] Geobytes Inc.: `http://www.geobytes.com/`.

- **FreeLing Measure Recognizer and Classifier**: a module for a fine-grained classification of measures has been created. This module was added to Freeling and recognises the following measure classes: *acceleration*, *density*, *digital*, *dimension*, *energy*, *extent*, *flow*, *frequency*, *power*, *pressure*, *size*, *speed*, *temperature*, *time*, and *weight*.
- **Temporal expressions grammar**: this process recognises complex temporal expressions both in the questions and in the passages. It is a recogniser based on a grammar of temporal expressions (composed by 73 rules) which detects four types of such expressions:
  * *Date*: A specific day, including day, day of the week (most times calculated), month and year (and eventually the time).
  * *Date_range*: Period of time, spanning between two specific dates or expressions such as "in 1910" (which would be equivalent to the period between January 1st 1910 and December 31st 1910), but also the seasons or other well-known periods of the year.
  * *Date_previous*: the period previous to a specific date.
  * *Date_after*: the period subsequent to a specific date.

  Moreover, in all the four types, not only absolute dates or periods are detected, but also dates relative to the current date, in expressions such as "el próximo viernes" (next Friday),"ayer" (yesterday), or "a partir de mañana" (from tomorrow on). These relative dates are converted into absolute according to the date of the document in which they are found.

The application of the language dependent linguistic resources and tools to the text of the question is represented in two structures:

- **Sent**, which provides lexical information for each word: form, lemma, POS tag (Eagles tagset), semantic class of NE, list of EWN synsets and, finally, whenever possible the verbs associated with the actor and the relations between locations and their nationality.
- **Sint**, composed by two lists, one recording the syntactic constituent structure of the question (basically nominal, prepositional and verbal phrases) and the other collecting the information of dependencies and other relations between these components.

– **Question Refinement**. This module contains two components: a tokenizer and a parser (processing the lexical structure of Question Pre-processing step). The tokenizer refines and sometimes modifies the sent structure. Basically the changes can affect the NEs occurring in the question and their local context (both the segmentation and the classification can be affected). Taking evidences from the local context a NE can be refined (e.g. its label can change from location to city), reclassified (e.g. passing from location to organization), merged with another NE, etc. Most of the work of the tokenizer relies on a set of trigger words associated to NE types, especially locations. We have collected this set from the Albayzin corpus (a corpus of about 6,887 question patterns in Spanish on Spain's geography domain, [4]). The parser uses a DCG grammar learned from the Albayzin corpus and tuned with the CLEF 2004 questions. In addition of triggers, the grammar uses a set of

introducers, patterns of lemmas as "dónde" (where), "qué ciudad" (which city), etc. also collected from Albayzin corpus.

- **Environment Building.** The semantic process starts with the extraction of the semantic relations that hold between the different components identified in the question text. These relations are organized into an ontology of about 100 semantic classes and 25 relations (mostly binary) between them. Both classes and relations are related by taxonomic links. The ontology tries to reflect what is needed for an appropriate representation of the semantic environment of the question (and the expected answer). The environment of the question is obtained from *Sint* and the information included in *Sent*. A set of about 150 rules was built to perform this task. Only minor changes have been performed in this module, so refer to [5] for details.
- **Question Classification**. This component uses 72 hand made rules to extract the Question Type (QT). These rules use a set of introducers (e.g. 'where'), and the predicates extracted from the environment (e.g. location, state, action,...) to detect the QT (currently, 25 types). The QT is needed by the system when searching the answer. The QT focuses the type of expected answer and provides additional constraints.
- **Semantic Constraints Extraction.** Depending on the QT, a subset of useful items of the environment has to be selected in order to extract the answer. Sometimes additional relations, not present in the environment, are used and sometimes the relations extracted from the environment are extended, refined or modified. We define in this way the set of relations (the semantic constraints) that are supposed to be found in the answer. These relations are classified as mandatory, Mandatory Constraints (MC), (i.e. they have to be satisfied in the passage) or optional, Optional Constraints (OC), (if satisfied the score of the answer is higher). In order to build the semantic constraints for each question a set of rules (typically 1 or 2 for each type of question) has been manually built. Although the structure of this module has not changed from our CLEF 2004 system, some of the rules have been modified and additional rules have been included for taking profit of the richer information available for producing more accurate Semantic Constraints (a set of 88 rules is used).

### 2.3 Passage Retrieval

The Passage Retrieval subsystem is performed using the *Lucene* Information Retrieval system. The PR algorithm uses a data-driven query relaxation technique: if too few passages are retrieved, the query is relaxed first by increasing the accepted keyword proximity and then by discarding the keywords with the lowest priority. The reverse happens when too many passages are extracted. Each keyword is assigned a priority using a series of heuristics fairly similar to [9]. The Passage Retrieval subsystem has been improved with the following components:

- **Temporal Constraints Keywords Search**: when a keyword is a temporal expression, the PR system returns passages that have a temporal expression that satisfies the constraint detected by our temporal grammar.

– **Coreference resolution**: to enhance the recall in the Answer Extraction modules, we apply a coreference resolution algorithm to the retrieved passages. We use an adaptation of the limited-knowledge algorithm proposed in [10]. We start by clustering the Named Entities in every passage according to the similarity of their forms (trying to capture phenomena as acronyms). For Named Entities classified as Person we use a first name gazetteer[4] to classify them as masculine or feminine. By the clustering procedure we get the gender information for the occurrences of the name where the first name does not appear. After that, we detect the omitted pronouns and the clause boundaries using the method explained in [7], and then apply the criteria of [10] to find the antecedent of reflexive, demostrative, personal and omitted pronouns among the noun phrases in the 4 previous clauses.

### 2.4 Factoid Answer Extraction

After PR, for factoid AE, two tasks are performed in sequence: Candidate Extraction (CE) and Answer Selection (AS). In the first component, all the candidate answers are extracted from the highest scoring sentences of the selected passages. In the second component the best answer is chosen.

– **Candidate Extraction**. The answer extraction process is carried out on the set of passages obtained from the previous subsystem. These passages are segmented into sentences and each sentence is scored according to its semantic content (see [8]). The linguistic process of extraction is similar to the process carried out on questions and leads to the construction of the environment of each candidate sentence. The rest is a mapping between the semantic relations contained in this environment and the semantic constraints extracted from the question. The mandatory restrictions must be satisfied for the sentence to be taken into consideration; the satisfaction of the optional constraints simply increases the score of the candidate. The final extraction process is carried out on the sentences satisfying this filter. The knowledge source used for this process is a set of extraction rules with a credibility score. Each QT has its own subset of extraction rules that leads to the selection of the answer. The application of the rules follows an iterative approach. In the first iteration all the semantic constraints must be satisfied by at least one of the candidate sentences. If no sentence has satisfied the constraints, the set of semantic constraints is relaxed by means of structural or semantic relaxation rules, using the semantic ontology. Two kinds of relaxation are considered: i) moving some constraint from MC to OC and ii) relaxing some constraint in MC substituting it for another more general in the taxonomy. If no candidate sentence occurs when all possible relaxations have been performed the question is assumed to have no answer.

[4] By Mark Kantrowitz, `http://www-2.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/nlp/corpora/names`

– **Answer selection**. In order to select the answer from the set of candidates, the following scores are computed for each candidate sentence: i) the rule score (which uses factors such as the confidence of the rule used, the relevance of the OC satisfied in the matching, and the similarity between NEs occurring in the candidate sentence and the question), ii) the passage score, iii) the semantic score (defined previously) , iv) the relaxation score (which takes into account the level of rule relaxation in which the candidate has been extracted). For each candidate the values of these scores are normalized and accumulated in a global score. The answer to the question is the candidate with the best global score.

## 3 Definitional QA System

The Definitional QA System has three phases: Passage Retrieval, Sentence Extraction, and Sentence Selection. In the first phase, an index of documents has been created using Lucene. The search index has two fields: one with the lemmas of all non-stop words in the documents, and another with the lemmas of all the words of the documents that begin with a capital letter. The target to define is lemmatized, stopwords are removed and the remaining lemmas are used to search into the index of documents. Moreover, the words of the target that begin with a capital letter are lemmatized; the final query sent to Lucene is a complex one, composed of one sub-query using document lemmas and another query containing only the lemmas of the words that begin with a capital letter. This second query is intended to search correctly the targets that, although being proper names, are composed or contain common words. For example, if the target is "Sendero Luminoso", documents containing the words "sendero" or "luminoso" as common names are not of interest; the occurrence of these words is only of interest if they are proper names, and as a simplification this is substituted by the case the words begin with a capital letter. The score of a document is the score given by Lucene. Once selected a number of documents (50 in the current configuration), the passages (blocks of 200 words) that refer to the target are selected for the next phase.

The objective of the second phase is to obtain a set of candidate sentences that might contain the definition of the target. As definitions usually have certain structure, as appositions or copulative sentences, a set of patterns has been manually developed in order to detect these and other expressions usually associated with definitions (for example, <phrase> , <target>, or <phrase> "ser" <target>). The sentences that match any of these patterns are extracted.

In the last step, one of the sentences previously obtained has to be given as the answer. In order to select the most likely sentence, an assumption has been made, in the sense that the words most frequently co-occurring with the target will belong to its definition. Thus, the frequency of the words (strictly, their lemmas) in the set of candidate sentences is computed and the sentence given as answer is the one whose words sum up a higher value of relative frequency.

## 4 Results

This section evaluates the behaviour of our system in CLEF 2005. We evaluated the three main components of our Factual QA system and the global results:

- **Question Processing**. This subsystem has been manually evaluated for factoid questions (see Table 1) in the following components: POS-tagging, NER and NE Classification (NEC) and QC. These results are accumulatives.

**Table 1.** Results of Question Processing evaluation.

| Question Type | Subsystem | Total units | Correct | Incorrect | Accuracy | Error |
|---|---|---|---|---|---|---|
| FACTOID | POS-tagging | 1122 | 1118 | 4 | 99.64% | 0.36% |
| | NE Recognition | 132 | 129 | 3 | 97.73% | 2.27% |
| | NE Classification | 132 | 87 | 45 | 65.91% | 34.09% |
| | Q. Classification | 118 | 78 | 40 | 66.10% | 33.89% |
| TEMPORAL | POS-tagging | 403 | 402 | 1 | 99.75% | 0.25% |
| | NE Recognition | 64 | 56 | 8 | 87.50% | 12.50% |
| | NE Classification | 64 | 53 | 11 | 82.81% | 17.19% |
| | Q. Classification | 32 | 27 | 5 | 84.37% | 15.62% |

- **Passage Retrieval**. This subsystem was evaluated using the set of correct answers given by the CLEF organization (see Table 2). We computed two measures: the first one (called *answer*) is the accuracy taking into account the questions that have a correct answer in its set of passages. The second one (called *answer+docID*) is the accuracy taking into account the questions that have a minimum of one passage with a correct answer and a correct document identifier in its set of passages. For factoid questions the two runs submitted differ in the parameters of the passage retrieval module: i) the maximum number of documents retrieved was 1200 (run1) and 1000 (run2), ii) the windows proximity was: (run1: 60 to 240 lemmas; run2: 80 to 220 lemmas), iii) the threshold for minimum passages: 4 (run1) and 1 (run2), iv) the maximum number of passages retrieved: 300 (run1) and 50 (run2).

**Table 2.** Passage Retrieval results (accuracy).

| Question type | Measure | run1 | run2 |
|---|---|---|---|
| FACTOID | Acc. (*answer*) | 78.09% (82/105) | 76.19% (80/105) |
| | Acc. (*answer+docID*) | 64.76% (68/105) | 59.05% (62/105) |
| TEMPORAL | Acc. (*answer*) | 50.00% (13/26) | 46.15% (12/26) |
| | Acc. (*answer+docID*) | 34.61% (9/26) | 30.77% (8/26) |

– **Answer Extraction**. The evaluation of this subsystem (see Table 3) uses the *answer+docID* and *answer* accuracies described previously.

**Table 3.** Factoid Answer Extraction results (accuracy).

| Question Type | Accuracy Type | run1 | run2 |
|---|---|---|---|
| FACTOID | Acc. (*answer*) | 29.27% (24/82) | 26.25% (21/80) |
| | Acc. (*answer+docID*) | 35.29% (24/68) | 33.87% (21/62) |
| TEMPORAL | Acc. (*answer*) | 15.38% (2/13) | 33.33% (4/12) |
| | Acc. (*answer+docID*) | 22.22% (2/9) | 50.00% (4/8) |

– **Global Results**. The overall results of our participation in CLEF 2005 Spanish monolingual QA task are listed in Table 4.

**Table 4.** Results of TALP-QA system at CLEF 2005 Spanish monolingual QA task.

| Measure | run1 | run2 |
|---|---|---|
| Total Num. Answers | 200 | 200 |
| Right | 58 | 54 |
| Wrong | 122 | 133 |
| IneXact | 20 | 13 |
| Unsupported | 0 | 0 |
| Overall accuracy | 29.00% (58/200) | 27.00% (54/200) |
| Accuracy over Factoid | 27.97% (33/118) | 25.42% (30/118) |
| Accuracy over Definition | 36.00% (18/50) | 32.00% (16/50) |
| Accuracy over Temporal Factoid | 21.88% (7/32) | 25.00% (8/32) |
| Answer-string "NIL" returned correctly | 25.92% (14/54) | 22.41% (13/58) |
| Confidence-weighted Score | 0.08935 (17.869/200) | 0.07889 (15.777/200) |

## 5 Evaluation and Conclusions

This paper summarizes our participation in the CLEF 2005 Spanish monolingual QA evaluation task. Out of 200 questions, our system provided the correct answer to 58 questions in run1 and 54 in run2. Hence, the global accuracy of our system was 29% and 27% for run1 and run2 respectively. In comparison with the results of the last evaluation (CLEF 2004), our system has reached a little improvement (24% and 26% of accuracy). Otherwise, we had 20 answers considered as inexact. We think that with a more accurate extraction phase we could extract correctly more questions and reach easily an accuracy of 39% . We conclude with a summary of the system behaviour for the three question classes:

- **Factoid questions**. The accuracy over factoid questions is 27.97% (run1) and 25.42% (run2). Although no direct comparison can be done using another test collection, we think that we have improved slightly our factoid QA system with respect to the results of the CLEF 2004 QA evaluation (18.89% and 21.11%) in Spanish. In comparison with the other participants of the CLEF 2005 Spanish QA track, our system has obtained good results in the following type of questions: location and time. On the other hand, our system has obtained a poor performance in the classes: measure and other.

  - **Question Processing**. In this subsystem the Question Classification component has an accuracy of 66.10%. This result means that there is no great improvement with respect to the classifier used in CLEF 2004 (it reached a 58% of accuracy). These values are influenced by the previous errors in the POS, NER and NEC subsystems. On the other hand, NEC errors have increased substantially with respect to the previous evaluation. NEC component achieved an error rate of 34.09%.
  - **Passage Retrieval**. We evaluated that 78.09% (run1) and 76.19% (run2) of questions have a correct answer in their passages. Taking into account the document identifiers the evaluation shows that 64.76% (run1) and 59.05% (run2) of the questions are really supported. This subsystem has improved substantially its results in comparison with the CLEF 2004 evaluation (48.12% and 43.12% of *answer+docID* accuracy).
  - **Answer Extraction**. The accuracy of the AE module for factoid questions for which the answer and document identifier occurred in our selected passages was of 35.29% (run1) and 33.87% (run2). This means that we have not achieved an improvement of our AE module, since the results for this part in CLEF 2004 were 23.32% (run1) and 28.42% (run2), evaluated only with answer accuracy. This is the subsystem that performs worst and needs a substantial improvement and tuning.

- **Definition questions**. This subsystem has reached a performance of 36% (run1) and 32% (run2) of right answers. The difference between the two runs lies in the different priority values assigned to each definitional pattern. The system has failed mainly in giving exact answers. The main cause of error has been the failure to correctly extract the exact sentence defining the target, as in 15 questions there were more words than just the definition, and thus the answer was marked as inexact. Otherwise, 33 questions would have had a right answer, and thus a 66% performance would have been achieved.
- **Temporal Factoid Questions**. The accuracy over temporal factoid questions is 21.88% (run1) and 25.00% (run2). We detected poor results in the PR subsystem. The accuracy of PR with answer and document identifiers is 34.61% (run1) and 30.77% (run2). These results are due to the fact that some questions are temporally restricted by event. These questions have nested questions and we processed these questions as one unique question.

## Acknowledgements

This work has been partially supported by the European Commission (CHIL, IST-2004-506909), the Spanish Research Department (ALIADO, TIC2002-04447-C02), the Ministry of Universities, Research and Information Society (DURSI) of the Catalan Government, and the European Social Fund. Daniel Ferrés is supported by a UPC-Recerca grant from Universitat Politècnica de Catalunya (UPC). TALP Research Center, is recognized as a Quality Research Group (2001 SGR 00254) by DURSI. The authors would like to express their gratitude in particular to Lluís Padró and Mihai Surdeanu.

## References

1. J. Atserias, J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Márquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. Morphosyntactic Analisys and Parsing of Unrestricted Spanish Text. In *Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC*, pages 603–610, Granada, Spain, May 1998.
2. Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC*, Lisbon, Portugal, 2004.
3. Xavier Carreras, Lluís Màrquez, and Lluís Padró. Named Entity Extraction using AdaBoost. In *Proceedings of CoNLL-2002*, pages 167–170. Taipei, Taiwan, 2002.
4. J. Diaz, A. Rubio, A. Peinado, E. Segarra, N. Prieto, and F. Casacuberta. Development of Task-Oriented Spanish Speech Corpora. In *Procceedings of the First International Conference on Language Resources and Evaluation*, pages 497–501, Granada, Spain, May 1998. ELDA.
5. Daniel Ferrés, Samir Kanaan, Alicia Ageno, Edgar González, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo. The TALP-QA System for Spanish at CLEF 2004: Structural and Hierarchical Relaxing of Semantic Constraints. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *CLEF*, volume 3491 of *Lecture Notes in Computer Science*, pages 557–568. Springer, 2004.
6. Daniel Ferrés, Samir Kanaan, Edgar González, Alicia Ageno, Horacio Rodríguez, Mihai Surdeanu, and Jordi Turmo. TALP-QA System at TREC 2004: Structural and Hierarchical Relaxation Over Semantic Constraints. In *Proceedings of the Text Retrieval Conference (TREC-2004)*, 2005.
7. A. Ferrández and J. Peral. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, (ACL'2000)*, 2000.
8. Marc Massot, Horacio Rodríguez, and Daniel Ferrés. QA UdG-UPC System at TREC-12. In *Proceedings of the Text Retrieval Conference (TREC-2003)*, pages 762–771, 2003.
9. D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Gîrju, and V. Rus. LASSO: A tool for surfing the answer net. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, 1999.
10. M. Saiz. *Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español.* PhD thesis, Universidad de Alicante, 2002.