The TALP-QA System for Spanish at CLEF-2005

Daniel Ferrés, Samir Kanaan, Edgar González, Alicia Ageno, Horacio Rodríguez, and Jordi Turmo
TALP Research Center
Software Department
Universitat Politècnica de Catalunya
{*dferres,skanaan,egonzalez,ageno,horacio,turmo*}*@lsi.upc.edu*

Abstract

This paper describes TALP-QA, a multilingual open-domain Question Answering (QA) system that processes both factoid (normal and temporally restricted) and definition questions. The system is described and evaluated in the context of our participation in the CLEF 2005 Spanish Monolingual QA task.

Factoid Questions are processed using a typical QA architecture: Question Processing (QP), Passage Retrieval (PR) and Answer Extraction. Our approach to extract answers from factoid questions is to build a semantic representation of the questions and the sentences in the passages retrieved for each question. A set of Semantic Constraints (SC) are extracted for each question. An answer extraction algorithm extracts and ranks sentences that satisfy the SCs of the question. If matches are not possible the algorithm relaxes the SCs structurally (removing constraints or making them optional) and/or hierarchically (generalizing the constraints using a taxonomy).

The answers to definition questions are retrieved in three steps: first, the 50 most relevant documents with respect to the target to define are retrieved, from which the passages referring to the target are retrieved; second, candidate sentences referring to the target are extracted from the previous set of documents, and last a defining sentence is selected. This selection is done using a score computed with the frequency of the words that appear in the candidate sentences.

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, Factoid questions, Definition questions, Temporally restricted questions

1 Introduction

This paper describes TALP-QA, a multilingual open-domain Question Answering (QA) system under development at UPC for the past 3 years. The paper focuses on our participation in the CLEF 2005 evaluation (i.e. most of the paper describes with some details the improvements over the previous system that have been included for this evaluation). Our aim in developing TALP-QA has been to build a system as far as possible language independent, where language dependent modules could be substituted to allow the system to be applied to different languages. 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].

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 the general architecture. 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 System Description

2.1 Overview

The system architecture has changed with respect to the architecture used in CLEF 2004. In those architecture, the Definitional QA system used the Question Processing (QP) and Passage Retrieval (PR) modules created for the Factual QA system. Actually, factual and definition questions are treated with two different architectures. Definitional QA architecture has been build with specific Question Processing and Passage Retrieval modules for these kind of questions.

On the other hand, Factual QA system has been modified in order to deal with temporally restricted questions. These modifications affected specially the Question Processing and the Passage Retrieval modules. Finally, the Question Classification module has been changed.

2.2 Factual QA System

The system architecture for factual questions follows the most commonly used schema, splitting the process into three subsystems that are performed sequentially (as shown in Figure 1): Question Processing, Passage Retrieval and Answer Extraction (AE). The QA components may contain iterative algorithms (e.g. Passage Retrieval) but no feedback is propagated to the previous modules.

These subsystems are described below, but first we will describe some pre-processing tasks that were carried out on the document collection (the EFE corpus in this case). As mentioned, our aim is to develop a language independent system. Language dependent components are only included in the Question Pre-processing and Passage Pre-processing components, and can be substituted by components for other languages.

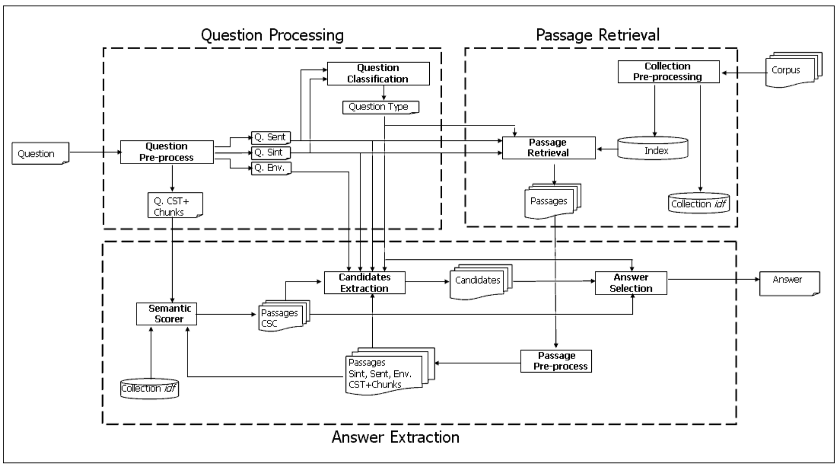

Figure 1: Architecture of TALP-QA Factual system.

2.2.1 Collection Pre-processing

We have used the *Lucene*[1] Information Retrieval (IR) engine to perform the PR task. Before CLEF 2005 we indexed the entire EFE collection: EFE 1994 and EFE 1995 (i.e. 454,045 documents). We pre-processed the whole collection with linguistic tools (described in [5]) to mark the part-of-speech (POS) tags, lemmas and Named Entities (NE). After this process the collection is parsed with a partial parser (described in the next sub-section). This information was used to built an index that contains the following parts for each document:

- Lemmatized and NERC text with syntactic information: this part is built using the lemmas of the words, POS and the results of the Named Entity Recognition and Classification (NERC) module. The text indexed also contains the syntactic information providing from the parsing step. This text is then indexed and used in the PR module.
- Original text with NE recognition with syntactic information: the original text that is retrieved when a query succeeds on the lemmatized text. This text contains also the syntactic information provided by the parser.

As an additional knowledge source that will be used in the AE task, an *idf* weight is computed at document level for the whole collection.

2.2.2 Question Processing

The main goal of this subsystem is to detect the expected answer type and to generate the information needed for the other subsystems. For PR, the information needed is basically lexical (POS, lemmas and NERC classes) and syntactic, and for AE, lexical, syntactic and semantic. We use a language-independent formalism to represent this information. We use the same semantic primitives and relations for both languages (English and Spanish) processed by our system.

A key point in QP is the Question Classification (QC) subtask. Without an accurate QC it is very difficult (often impossible) to get an appropriate answer. The results of QC in our previous attempt (in CLEF 2004) were rather low (only 58.33% accuracy) and we decided to proceed to an in depth revision of this component. As was explained in [5] the low accuracy obtained is basically due to two facts:

- The dependence on errors of previous tasks, namely POS-tagging (2.28% errors), NER (4.37%) and NEC (25.14%) that are accumulatives.
- The use for classification of a ILP learner that has learned 25 binary classifiers. The learners was trained with the manual translation of questions from TREC8 and TREC9 (about 900 questions) that were used for the same task in TREC. 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.

As reported in [5] the result of the semantic processing subtask was also low (45% accuracy on environment building, 42.78% on mandatory constraints and 72.78% on optional constraints). Although the consequences of these errors are not usually fatal as in the case of QC we tried to face both problems in the new QP module.

For facing these objectives, the QP module is structured into five components, We will describe next these components focussing in those having changed from our previous system (see [5] for details):

[1] http://jakarta.apache.org/lucene

- **Question Pre-processing**. This subsystem is basically the same component of our previous system with some improvements (described below). 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* (described in [7] and [5]). These tools are used for the linguistic processing of both the questions and the passages. The main improvements on these tools refer to:
  - **Geographical gazetteers**: the use of geographic gazetteers for increasing the accuracy in the NEC task. 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). The gazetteers used were:
    * GEOnet Names Server (GNS)[2]: we used the non-ambiguous places (a subset of the most important classes was used: cities, rivers, mountains, states, countries, regions,..) extracted from the GNS gazetteer. This gazetteer covers worldwide excluding the United States and Antarctica, with 5.3 million entries. A total of 126,941 non-ambiguous locatives were extracted from this gazetteer.
    * *GeoWorldMap*[3] *World Gazetteer*: a gazetteer with approximately 40,594 entries of the most important countries, regions and cities of the world.
    * *Albayzin Gazetteer*: a gazetteer of 758 place names from the Spanish geography existing in the speech corpus Albayzin [4].
  - **FreeLing Measure Recognizer and Classifier**: in the CLEF 2004 evaluation we detected that our system achieved poor results in questions that have a measure as answer. For this reason, we decided to improve the resolution of this kind of questions by using a module with a fine-grained classification of measures. This module was added to Freeling and recognizes the following classes: *acceleration*, *density*, *digital*, *dimension*, *energy*, *extent*, *flow*, *frequency*, *power*, *pressure*, *size*, *speed*, *temperature*, *time*, and *weight*. A module to recognize measures was added to Freeling and some parts of this software were modified:
    * *Tokenizer*: a list of measure's abbreviations was added (e.g. *mm2*, *mms2*,..).
    * *Dictionary*: a list of abbreviations with their lemma and POS was added to the dictionary (e.g. *kms kilómetro NCMS000*).
    * *Multiwords*: we added a list multiwords related to measure units to the Freeling multiwords file (e.g. *kilómetros_quadrados*).
    * *Measure Patterns*: a file that relates the abbreviations, lemmas and multiwords to a measure unit class (e.g. *Volume_cl cl cl. centílitro centílitros*).
  - **Temporal expressions grammar**: in order to deal with the new category of questions of temporally restricted factoids, we have incorporated an additional process to the linguistic processing described above. This process is applied after Freeling, in order to recognise more complex temporal expressions both in the questions and in the passages. It is a recogniser based on a grammar of temporal expressions which detects four types of such expressions:
    1. *Date*: A specific day, including day, day of the week (most times calculated), month and year (and eventually the time).
    2. *Date_range*: Period of time, spanning between two specific dates.
    3. *Date_previous*: the period previous to a specific date.
    4. *Date_after*: the period subsequent to a specific date.

[2]**GNS**. http://gnswww.nima.mil/geonames/GNS/index.jsp

[3]Geobytes Inc.: Geoworldmap database containing cities, regions and countries of the world with geographical coordinates. http://www.geobytes.com/.

The grammar is composed by 73 rules: 18 rules for the detection of specific dates, 44 rules for the recognition of date ranges, 5 rules for previous dates and 6 for periods later to a certain date. In the second type, not only specific periods of time are detected, 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. 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), "la próxima primavera" (next spring), or "a partir de mañana" (from tomorrow on). This case will only be found in the passages, so these relative dates are converted into absolute according to the date of the document in which they are found.

The application of these language dependent linguistic resources and tools to the text of the question is represented in two structures, see the example in Figure 2:

- **Sent**, which provides lexical information for each word: form, lemma, POS tag (an Eagles compliant rich tagset was used), 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.

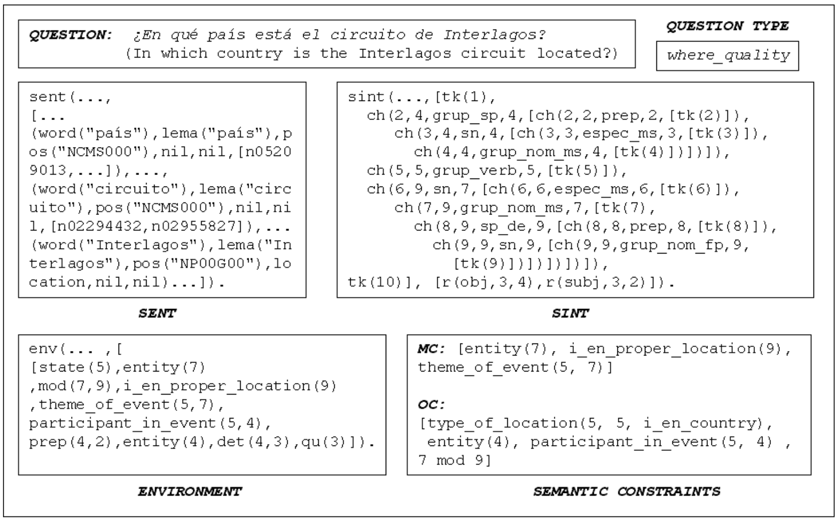

Figure 2: Results of pre-processing of a question.

- **Question Refinement**. While Question Pre-processing step consists in the application of a pipe of general purpose linguistic processors, this step has been designed specifically for questions. 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.

- **Question Classification**. This component takes profit of the results of previous ones. It is composed of a rule based system, using 72 hand made rules. See below tree rules related to the question class *where_quality*. These rules use a set of introducers (e.g. 'where'), and the predicates extracted from the environment (e.g. location, state,action,...) to classify the questions.

```
 classify_question('where',_,_,_,_,_,D2,_,where_quality,'23'):-
  \+ member('action':_,D2), member('prop_loc':_,D2).

classify_question('where',_,_,_,_,_,D2,D3,where_quality,'24'):-
  member('state':_,D2), \+ member('location':1,D3).

classify_question('where',_,_,_,_,D1,_,_,where_quality,'25'):-
  (member('location':_,D1); member('others':_,D1)).
```

  The QC rules are used to extract the Question Type (QT). The QT is the most important information we need to extract from the question text. The QT is needed by the system when searching the answer. Failure to identify the QT practically disables the correct extraction of the answer. Currently we are working with about 25 QTs. The QT focuses the type of expected answer and provides additional constraints. For instance, when the expected type of the answer is a person, two types of questions are considered, *Who_action*, which indicates that we are looking for a person who performs a certain action and *Who_person_quality*, that indicates that we are looking for a person having the desired quality. The action and the quality are the parameters of the corresponding QT. The following are examples of questions correctly classified respectively as *Who_person_quality* and *Who_action* type:

  - *¿Quién fue jefe del XII Gobierno de Israel? (Who was the head of the XII Israel government?)*
  - *¿Quién ganó el Premio Nobel de Literatura en 1994? (Who won the Nobel Prize for Literature in 1994?)*

- **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. The environment extracted from a question is presented in Figure 2.

- **Semantic Constraints Extraction.** The environment tries to represent the whole semantic content of the question. However, not all the items belonging to the environment are useful to extract the answer. So, depending on the QT, a subset of the environment has to be extracted. 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).

2.2.3 Passage Retrieval

The main function of the passage retrieval component is to extract small text passages that are likely to contain the correct answer. Document retrieval is performed using the *Lucene* Information Retrieval system. The passage retrieval 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 [10]. For example, a proper noun is assigned a higher priority than a common noun, the question focus word (e.g. "state" in the question "What state has the most Indians?") is assigned the lowest priority, and stop words are removed.

The Passage Retrieval subsystem has been improved with the following components:

- **Index with NERC and syntactic information**: the current index for CLEF 2005 has been build with the following information: forms, lemmas, POS, NERC tags and syntactic information. Indexing and Retrieving this information avoids pre-processing steps after the retrieval and the Passage Pre-processing module becomes faster.
- **Temporal Constraints Keywords Search**: our Passage Retrieval module was modified to deal with temporally restricted factoid questions. When a keyword is a temporal expression, the PR system returns passages that have a temporal expression that satisfies the temporal keyword constraint. The 4 types of temporal constrains are:
  1. *Date* : a specific date (e.g. "23 de Agosto de 1993" (*"23 August of 1993"*) is stored as [23/08/1993].
  2. *Date Range* : "in 1993" this constraint implies a date in the following range [1/1/1993-31/12/1993].
  3. *Date Previous*: this constraint means a date before a specific date.
  4. *Date After* : this constraint implies a date after a specific date.
- **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 [11].

  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 Entites 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 [8], and then apply the criteria of [11] to find the antecedent of reflexive, demostrative, personal and omitted pronouns among the noun phrases in the 4 previous clauses.

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

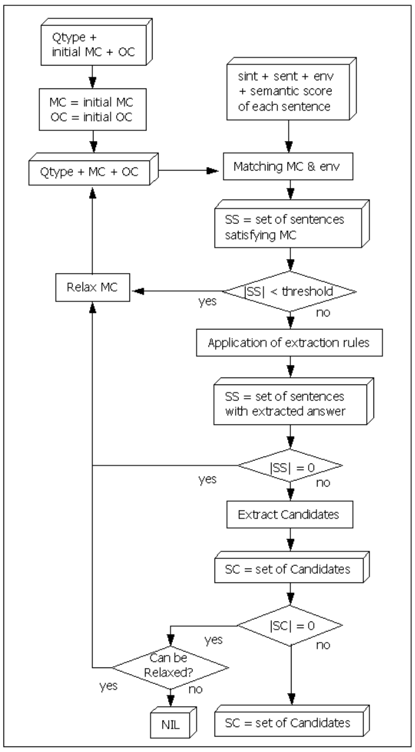

Figure 3: Candidates Extraction Relaxation Loop.

2.2.4 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 process of extraction of the answer is presented in Figure 3. This 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 using the $tf*idf$ weighting of the terms from the question and taxonomically related terms occurring in the sentence (see [9]).

  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 have to 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 constraint in the taxonomy. If no candidate sentence occurs when all possible relaxations have been performed the question is assumed to have no answer.

- **Answer selection**. In order to select the answer from the set of candidates, the following scores are computed for each candidate sentence:

  - 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.
  - The passage score, which uses the relevance of the passage containing the candidate.
  - The semantic score, defined previously.
  - 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.

2.3 Definition QA System

The Definition QA System has an architecture different from the Factual QA system (see Figure 4). The answers to definition questions are retrieved in three steps: first, the 50 most relevant documents with respect to the target to define are retrieved, from which the passages referring to the target are retrieved; second, sentences referring to the target are extracted from the previous set of documents, and last a defining sentence is selected, which will be the answer given by the system.

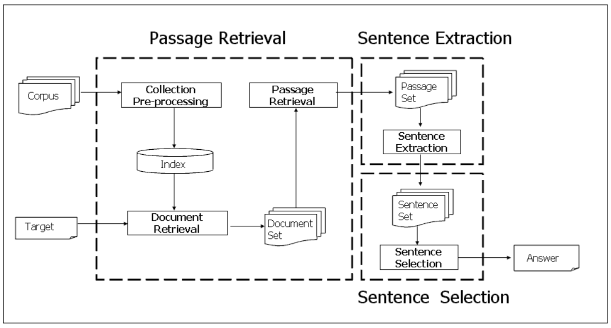

Figure 4: Architecture of TALP-QA Definitional system.

2.3.1 Document and Passage Retrieval

An index of documents has been created using Lucene that searches using lemmas instead of words. 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 subquery 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.

2.3.2 Sentence Extraction

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.

2.3.3 Sentence Selection

In the last step, one of the sentences previously obtained has to be given as the answer to the definition question. In order to select the most likely sentence, an assumption has been made, in the sense that the words most frequently co-ocurring 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.

3 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: basic NLP tools (POS, NER and NE Classification (NEC)) and Question Classification. 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**. The evaluation of this subsystem was performed using the set of correct answers given by the CLEF organization (see Table 2). We submitted two runs. These runs differ only in the parameters of the passage retrieval module for factoid questions:

– Number of documents retrieved: we have chosen a maximum of 1200 documents in run1 and 1000 documents in run2.

– Windows proximity: in run1 the proximity of the different windows that can compose a passage was lower than run2 (run1: 60 to 240 lemmas; run2: 80 to 220 lemmas).

– Threshold for minimum passages: the PR algorithm relaxes the query to obtain more passages if the number of extracted passages is lower than this threshold. These values are: 4 (run1) and 1 (run2) passages.

– Number of passages retrieved: we have chosen a maximum of 300 passages in run1 and 50 passages in run2.

Table 2: Passage Retrieval results.

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

In this part 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.

- **Answer Extraction**. The evaluation of this subsystem (see Table 3) for factoid questions has been done in three parts: evaluation of the Candidate Extraction (CE) module, evaluation of the Answer Selection (AS) module and finally evaluation of the AE subsystem's global accuracy for factoid questions in which the answer appears in our selected passages. This evaluation uses the *answer+docID* and *answer* accuracies described above.

Table 3: Factoid Answer Extraction results.

Question Type	Subsystem	Accuracy Type	run1	run2
FACTOID	Candidate Extraction	Acc. (*answer*)	40.24% (33/82)	37.5% (30/80)
		Acc. (*answer+docID*)	42.64% (29/68)	41.93% (26/62)
	Answer Selection	Acc.(*answer*)	72.72% (24/33)	70.00% (21/30)
		Acc. (*answer+docID*)	82.75% (24/29)	80.77% (21/26)
	Answer Extraction	Acc. (*answer*)	29.27% (24/82)	26.25% (21/80)
		Acc. (*answer+docID*)	35.29% (24/68)	33.87% (21/62)
TEMPORAL	Candidate Extraction	Acc. (*answer*)	15.38% (2/13)	33.33% (4/12)
		Acc. (*answer+docID*)	22.22% (2/9)	50.00% (4/8)
	Answer Selection	Acc. (*answer*)	100% (2/2)	100% (4/4)
		Acc. (*answer+docID*)	100% (2/2)	100% (4/4)
	Answer Extraction	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/Wrong	58/122	54/133
IneXact/Unsupported	20/0	13/0
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 correcty	25.92% (14/54)	22.41% (13/58)
Confidence-weighted Score	0.08935 (17.869/200)	0.07889 (15.777/200)

4 Evaluation and Conclusions

This paper summarizes our participation in the CLEF 2005 Spanish monolingual QA 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 an other 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**. In the PR 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 a 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.
- **Definition questions**. This subsystem has reached a performance of 36% right answers, and has failed mainly in giving exact answers. From a total set of 50 definition questions, 18

have been correctly answered by our system. 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. Clearly this is one aspect to improve in the current system.

- **Temporal Factoid Questions**. The accuracy over temporal factoid questions is 21.88% (run1) and 25.00% (run2). We detected poor results in the Passage Retrieval 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. These kind of questions must be splitted in two and processed separately.

Acknowledgements

This work has been partially supported by the European Commission (CHIL, IST-2004-506909) and the Spanish Research Dept. (ALIADO, TIC2002-04447-C02). TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government. 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. TALP-QA System for Spanish at CLEF-2004: Structural and Hierarchical Relaxing over Semantic Constraints. In C. Peters, P.D. Clough, G.J.F. Jones, J. Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*. Springer-Verlag LNCS, To appear, 2005.

[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] Daniel Ferrés, Marc Massot, Muntsa Padró, Horacio Rodríguez, and Jordi Turmo. Automatic Building Gazetteers of Co-referring Named Entities. In *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC*, Lisbon, Portugal, 2004.

[8] 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.

[9] 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.

[10] 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.

[11] 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.