# The role of lexical features in Question Answering for Spanish

Manuel Pérez-Coutiño, Manuel Montes-y-Gómez,
Aurelio López-López and Luis Villaseñor-Pineda

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Luis Enrique Erro No. 1, CP 72840, Sta. Ma. Tonantzintla, Pue., México.
{mapco,mmontesg,allopez,villasen}@inaoep.mx

**Abstract**. This paper describes the prototype developed in the Language Technologies Laboratory at INAOE for the Spanish monolingual QA evaluation task at CLEF 2005. The proposed approach copes with the QA task according to the type of question to solve (factoid or definition). In order to identify possible answers to factoid questions, the system applies a methodology centered in the use of lexical features. On the other hand, the system is supported by a pattern recognition method in order to identify answers to definition questions. The paper shows the methods applied at different stages of the system, with special emphasis on those used for answering factoid questions. Then the results achieved with this approach are discussed.

## 1 Introduction

Current information requirements call for efficient mechanisms capable of interaction with users in a natural way. Question Answering (QA) systems have been proposed as a feasible option for the creation of such mechanisms. Moreover, the research in this field shows a constant growth both in interest as well as in complexity [3]. This paper presents the prototype developed in the Language Technologies Laboratory at INAOE[1] for the Spanish monolingual QA evaluation task at CLEF 2005. The experiments performed this year by our group are a progression of our efforts reported last year [5] in the following aspects; a) the approach for answering factoid questions is centered in the analysis of the near context related to each named entity selected as candidate answer; b) the context used to discriminate candidate and final answers relies on the lexical information gathered by a shallow language processing (POS and named entities tagging) and statistical parameters. On the other hand, there are some important changes in the prototype architecture that allowed the system to have an improvement in performance (recall) at the initial stages of the QA task. At the same time, there have been some simplifications in the general architecture, which have allowed to get more control and flexibility in order to evaluate multiple system configurations and reduce error propagation from initial stages. For instance, we have

[1] http://ccc.inaoep.mx/labtl/

applied a shallow question classification process instead of a fine grain question classification; and the answer discrimination process relies only on the information located in the target documents, discarding internet searching and extraction modules of our previous prototype.

This paper is focused on the discussion of the proposed methodology for factoid question answering. Nevertheless, a section is presented with a brief description of the methods used for answering definition questions. The rest of this paper is organized as follows; section two describes the architecture of the prototype; from section three to section six the internal processes of the system are discussed; section seven discusses the results achieved by the system; and finally section eight contains our conclusions and discusses further work.

## 2 Prototype Architecture

As stated before, the system is based on the methodology proposed in the previous year [5] but with some significant modifications in the prototype. Figure 1 shows the main blocks of the system. Here the treatment of factoid and definition questions occurs separately.

Factoid questions resolution relies on a hybrid system involving the following stages: *question processing*, which includes the extraction of named entities and lexical context from the question, as well as question classification to define the semantic class of the answer expected to respond to a given question; *document processing*, where the preprocessing of the supporting document collection is done in parallel by a *passage retrieval system (PRS)* and a shallow NLP (similar to that performed in question processing); *searching*, where a set of candidate answers is gathered from a representation of the passages retrieved by the PRS; and finally *answer extraction*, where candidate answers are analyzed, weighted and ranked in order to produce the final answer recommendation of the system.

On the other hand, definition questions are treated directly with a method supported by a couple of lexical patterns that allow finding and selecting the set of possible answers. The following sections describe each of these stages.

## 3 Question Processing

QA systems traditionally perform a question processing stage in order to know in advance the semantic class of the answer expected for a given question and thus, reduce the searching space to only those information fragments related to instances of the semantic class previously determined. Our prototype implements this stage following a straightforward approach involving these steps:

1. Question is parsed with a set of heuristic rules in order to get its semantic class.
2. Question is tagged with the MACO POS tagger [1]
3. Named entities of the question are identified and classified using MACO.

The first step is responsible of identifying the semantic class of the expected answer. In the experiments performed with the training data set, we observed that

when the number of classes was minimal (just 3 classes: date, quantity and proper noun) it was possible to achieve similar results in precision to those achieved when we used a finer classification, for instance person, organization, location, quantity, date and other. Steps 2 and 3 produce information used later on, during searching to match questions and candidate answer context, contributing to the weighting scheme.

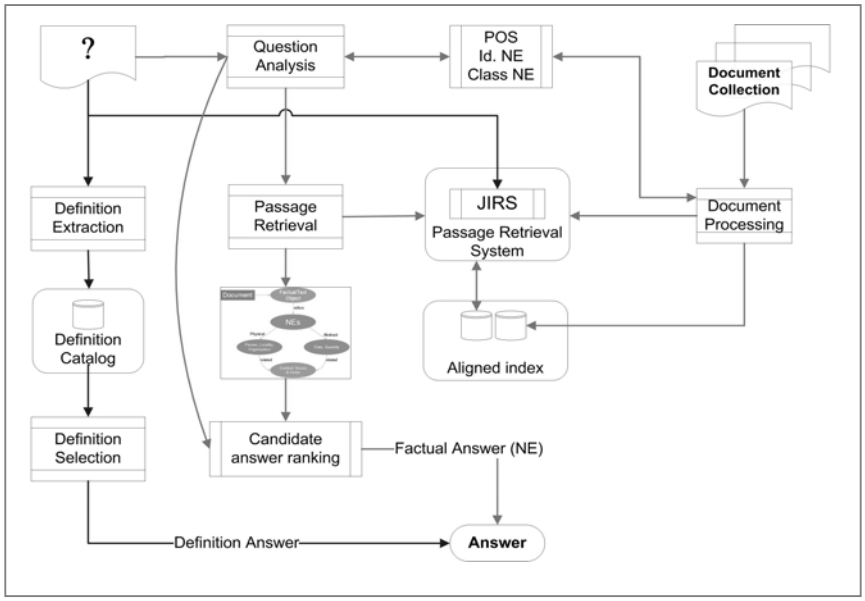


**Fig. 1.** Block diagram of the system. Factoid and definition questions are treated separately. Factoid questions require the following stages: question processing, document processing, searching, and answer selection. Definition questions use a pattern approach for definition extraction and selection processes

## 4 Document Processing

The prototype implements a hybrid approach for document processing that has allowed simplifying and increasing performance in this stage. The processing of target documents consists of two parts, first the whole document collection is tagged with MACO[1], gathering the POS tags as well as named entities identification and classification for each document in the collection. The second part of this stage is performed by the JIRS [2] passage retrieval system (PRS), that creates the index for the searching process. The index built by JIRS and the tagged collection are aligned phrase by phrase for each document in the collection. In this way, the system can retrieve later the relevant passages for a given question with JIRS, and then use their tagged form for the answer extraction process.

## 5 Searching

The searching stage is also performed in two steps. As we mentioned, the first step is to retrieve the relevant passages for the given question. This step is performed by JIRS, taking as input the question without any previous processing.

JIRS is a PSR specially suited for question answering. JIRS ranks the retrieved passages based on the computation of a weight for each passage. The weight of a passage is related to the size of the n-gram structure of the question that can be found in the passage. The larger the n-gram structure, the greater the weight assigned to the passage. The following example illustrates this concept.

Given the question "*Who is the president of Mexico?*", suppose that two passages returned the following text segments: "*Vicente Fox is the president of Mexico…*" ($p_1$) and "*The president of Spain visited Mexico in last February…*" ($p_2$).

The original question is divided into five sets of *n*-grams (5 is the number of question terms after removing the question word *Who*), these sets are the following:

**5-gram**: {"is the President of Mexico"}

**4-gram**: {"is the President of", "the President of Mexico"}

**3-gram**: {"is the President", "the President of", "President of Mexico"}

**2-gram**: {"is the", "the President", "President of", "of Mexico"}

**1-gram**: {"is", "the", "President", "of", "Mexico"}

Then, the five sets of *n*-grams from the two passages are gathered. The passage $p_1$ contains all the *n*-grams of the question (the 5-gram, the two 4-grams, the three 3-grams, the four 2-grams and the five 1-grams of the question). Therefore the similarity of the question with this passage is 1.

The sets of *n*-grams of the passage $p_2$ contain only the "*the President of*" 3-gram, the "*the President*"' and "*President of*" 2-grams and the following 1-grams: "*the*", "*President*", "*of*" and "*Mexico*". The similarity for this passage is lower than that for $p_1$ because the second passage is quite different with respect to the original question, although it contains all the relevant terms of the question.

A previous evaluation of JIRS [2] shows that the possible answer to a given question is found among the first 20 passages retrieved for over 60% of the training set.

Once the relevant passages are selected, the second step requires the POS tagged form of each passage in order to gather the representation used to extract the answer. Due to some technical constraints we were unable to finish the implementation for the alignment of the tagged collection and the JIRS index before test set release. Therefore the tagging of relevant passages was performed online with the annoyance of a couple extra hours for such processing.

Tagged passages are represented in the same way as proposed in [4] where each retrieved passage is modeled by the system as a factual text object whose content refers to several named entities[2] even when it could be focused on a central topic. The model assumes that the named entities are strongly related to their lexical context, especially to nouns (subjects) and verbs (actions). Thus, a passage can be seen as a set of entities and their lexical context. Such representation is used later in order to match the question representation against the set of best candidates gathered from passages.

[2] The semantic classes used rely on the capability of the named entity classifier, and could be one of these: persons, organizations, locations, dates, quantities, and miscellaneous.

## 6 Answer Extraction

### 6.1 Answering Factoid Questions

The system does not differentiate between simple and temporally restricted factoid questions in order to extract their possible answer. Given the set of retrieved passages and their representations (named entities and their contexts) the system computes a weight for each candidate answer (named entity) based on two main factors: a) the activation and deactivation of some features at different steps of the system, and b) the assigned weight computed with the formula 1.

The features listed in table 1 allow us to configure the system in order to change its behavior. For instance, deactivate the question classification step by allowing the final answer selection to rely only on statistical computations. The opposite case could be, deactivate frequency features and let the final answer selection to rely on the matching between question and candidate answers context.

$$\omega_A = \frac{t_q}{n} * \left( \frac{NE_q \cap NE_A}{|NE_q|} + \frac{C_q \cap C_A}{|C_q|} + \frac{F_A(P_i)}{F_A(P)} + \left(1 - \frac{P_i}{k-1}\right) \right) \quad (1)$$

i=1..k; k=number of passages retrieved by JIRS

Where $t_q$ is 1 if the semantic class of the candidate answer is the same as that of the question and 0 in other case; $n$ is a normalization factor based on the number of activated features, $NE$ is the set of named entities in the question ($q$) or in the candidate answer ($A$); $C$ is the context either for question ($q$) or candidate answer ($A$); $F_A(P_i)$ is the frequency of occurrence of the candidate answer in the passage $i$; $F_A(P)$ is the total frequency of occurrence of the candidate answer in the passages retrieved; and $1-(P_i/k-1)$ is an inverse relation for the passage ranking returned by JIRS.

**Table 1.** Features list used in factoid question answering

| Features | Function |
|---|---|
| 1. Question classification | Activate question classification step |
| 2. No. Classes | Defines the number of classes to use in question and named entity classification. |
| 3. Context elements | Define the elements included as part of a name entity context. They could be: named entities, common names, verbs, adjectives, adverbs, etc. |
| 4. Context length | Number of elements at left and right of a named entity to include in the context. |
| 5. Question Named Entities | Defines whether passages not containing named entities of the question are allowed. |
| 6. Context match | Intersection |
| 7. Frequency of occurrence | Number of times that a named entity appears as candidate answer in the same passage. |
| 8. JIRS ranking | Position of passage as retuned by JIRS. |
| 9. Passage length | Number of phrases in the passage retrieved. |

Once the system computes the weight for all candidate answers, these are ranked by decreasing order, taking as answer that with the greatest weight.

### 6.2 Answering Definitions

The method for answering definition questions exploits some regularities of language and some stylistic conventions of news notes to capture the possible answer for a given definition question. A similar approach was presented in [6,7].

The process of answering a definition question considers two main tasks. First, the definition extraction, which detects the text segments that contains the description or meaning of a term (in particular those related with the name of a person or an organization). Then, the definition selection, where the most relevant description of a given question term is identified and the final answer of the system is generated.

**6.2.1 Definition Extraction.** The language regularities and the stylistic conventions of news notes are captured by two basic lexical patterns. These patterns allow constructing two different definition catalogs. The first one includes a list of pairs of acronym-meaning. The second consists of a list of referent-description pairs.

In order to extract the acronym-meaning pairs we use an extraction pattern based on the use of parentheses.

$$w_1 \text{ <}meaning\text{>} \text{ ( <}acronym\text{> )} \qquad \textbf{(i)}$$

In this pattern, $w_1$ is a lowercase non-stop word, <*meaning*> is a sequence of words starting with an uppercase letter (that can also include some stop words), and <*acronym*> indicates a single word also starting with an uppercase letter.

By means of this pattern we could identify pairs like [*PARM – Partido Auténtico de la Revolución Mexicana*].

In contrast, the extraction of referent-description pairs is guided by the occurrence of a special kind of appositive phrases. This information was encapsulated in the following extraction pattern.

$$w_1\ w_2 \text{ <}description\text{>} \text{ , <}referent\text{> ,} \qquad \textbf{(ii)}$$

Where $w_1$ may represent any word, except a preposition, $w_2$ is a determiner, <*description*> is a free sequence of words, and <*referent*> indicates a sequence of words starting with an uppercase letter or appearing in the stop words list.

Applying this extraction pattern we could find pairs like [*Alain Lombard - El director de la Orquesta Nacional de Burdeos*].

**6.2.2 Definition Selection.** The main advantage of the extraction patterns is their generality. However, this generality causes the patterns to often extract non relevant information, i.e., information that does not indicate a relation acronym-meaning or concept-description.

Given that the catalogs contains a mixture of correct and incorrect relation pairs, it is necessary to do an additional process in order to select the most likely answer for a given definition question. The proposed approach is supported by the idea that, on one

hand, the correct information is more abundant than the incorrect, and on the other, that the correct information is redundant.

Thus, the process of definition selection considers the following two criteria:

1. The more frequent definition in the catalog has the highest probability to be the correct answer.
2. The largest and therefore more specific definitions tend to be the more pertinent answers.

The following example illustrates the process. Assuming that the user question is "*who is Félix Ormazabal?*", and that the definition catalog contains the records showed below. Then, the method selects the description "*diputado general de Alava*" as the most likely answer.

*Félix Ormazabal: Joseba Egibar:*
*Félix Ormazabal: candidato alavés:*
*Félix Ormazabal: diputación de este territorio:*
*Félix Ormazabal: presidente del PNV de Alava y candidato a diputado general:*
*Félix Ormazabal: nuevo diputado general*
*Félix Ormazabal: diputado Foral de Alava*
*Félix Ormazabal: través de su presidente en Alava*
*Félix Ormazaba : diputado general de Alava*
*Félix Ormazabal: diputado general de Alava*
*Félix Ormazabal: diputado general de Alava*

## 7 Experiments and Results

This section discusses some training experiments and the decision criteria used to select the configuration of the experiments evaluated at QA@CLEF2005 monolingual track for Spanish. Given that we have used the same modules for answering definition questions in all our runs for monolingual QA, including those described in "A Full Data-Driven System for Multiple Language Question Answering" (also in this volume), the discussion on these results and some samples have been documented in that paper. The rest of this document is intended to discuss the results on factoid question answering.

### 7.1 Training Experiments

As we mentioned earlier, the approach used in our system is similar to that used in [5], an analysis of such system showed that it was necessary to experiment with different values for the parameters involved in the answer extraction stage (see table 1). For instance, in [5] the system relied on a document model considering only nouns or verbs at left and right of named entities, within a lexical context of four elements. In order to improve our approach we performed several experiments using context lengths from four elements to the whole passage retrieved. We also tested different elements for the lexical context: i.e. nouns, proper nouns, verbs, adjectives and adverbs. Table 2 shows some configurations tested with the training set. Then, figure

2 shows the results achieved with the training set applying the configurations showed in table 2. Notice that these results correspond to the factoid question answering.

**Table 2.** Configurations of some experiments performed with the training set. First column refers to the features listed in table 1

| | **Exp. 1** | **Exp. 2** | **Exp. 3** | **Exp. 4** | **Exp. 5** | **Exp. 6** | **Exp. 7** | **Exp. 8** | **Exp. 9** |
|---|---|---|---|---|---|---|---|---|---|
| **1** | No | Yes | Yes | No | Yes | No | Yes | No | Yes |
| **2** | 0 | D,Q,NP | D,Q,P,O,G | 0 | D,Q,NP | 0 | D,Q,NP | 0 | D,Q,NP |
| **3** | V,NC,NE | V,NC,NE | V,NC,NE | V,NC,NE | V,NC,NE | V,NC,NE,QA | V,NC,NE,QA | V,NC,NE,QA | V,NC,NE,QA |
| **4** | 4 | 4 | 4 | 4 | 4 | 8 | 8 | Passage | Passage |
| **5** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **6** | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **7** | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **8** | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **9** | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 |

Figure 2 shows that the best performance was achieved with the "Exp. 7" which combines the following feature values, first the system classifies the question as one of the following classes: Date, Question, Proper Noun (which includes person, organizations and locations); next the system retrieves the relevant passages with length = 1 phrase, and builds the proper representation for each named entity found in it. At this stage, the context is formed by 8 elements at the left and right of the named entity and considers verbs, common names, named entities and adjectives. The extraction stage filters those candidate answers whose context does not contain any of the question named entity, and finally computes the weight for each candidates according to formula 1 (see table 2 for exp. 7 configuration).

Another interesting experiment was the analysis of the questions answered by this method. We estimate that the "union" of the results gathered with the configurations showed in table 2 could reach over 24% if the best configuration was selected online, i.e., for each question select the best configuration of the system which could return an accurate answer.

### 7.2 Evaluation

We participated in the evaluation with two runs, both were executed using the same configuration of experiment 7 (see table 2). The first one (inao051eses) analyzes the first 800 passages retrieved by JIRS, while our second run (inao052eses) analyzes only the first 100 passages retrieved by JIRS. Table 3 shows the results of the evaluation.

Despite the fact that our results (for factoid questions) were over 10% better than last year and one of the best for temporally restricted factoid questions, we believe that the approach described is close to its accuracy limit. The methodology is best suited for questions whose answer is commonly found in the near context of some reformulation of the question into the passages, while for other, more elaborated factoid questions, it is unable to identify the right answer. That is the case of questions whose expected answer is an object or some entity which can not be identified *a priori* by the shallow NLP used or without a knowledge base.

Another point to note is that in some cases, the statistical factor given by the frequency of occurrence of a candidate answer becomes a secondary aspect that could lead to a wrong selection of an answer.

We have begun some experiments with machine learning techniques in order to learn the appropriate system configuration based on the question attributes. Another direction in our research is to include more features that allow us to perform an improved selection and discrimination of candidate answers, moreover, that allow to consider objects and more entities that are currently excluded by the methodology.

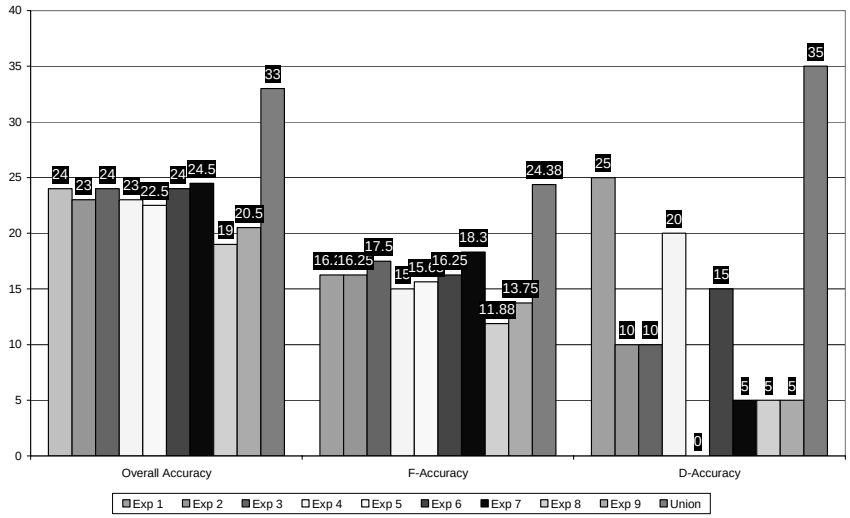


**Fig. 2.** Results achieved with training set, applying the configurations showed in table 2

**Table 3.** Results of submitted runs

| **Run** | ***inao051eses*** | ***inao052eses*** |
|---|---|---|
| Right | 84 (34F + 40D + 10 TRF) | 79 (32F + 40D + 7 TRF) |
| Wrong | 110 | 116 |
| ineXact | 5 | 4 |
| Unsupported | 1 | 1 |
| Overall Accuracy | 42.00% | 39.50% |
| Factoid Questions | 28.81% | 27.12% |
| Definition Questions | 80.00% | 80.00% |
| Temporally Restricted Factoid Questions | 31.25% | 21.88% |
| Answer string "NIL" | Precision= 0.23<br>Recall=0.80<br>F-score=0.36 | Precision= 0.19<br>Recall=0.80<br>F-score=0.31 |

## 8 Conclusions

This paper has presented an approach for QA in Spanish centered on the use of lexical features for factoid question resolution that is complemented with a pattern matching approach for definition question resolution. The results achieved in the monolingual track for Spanish have improved compared to our previous year performance by over 10% on factoid questions and over 30% on definition questions. It is important to note that the approach was able to answer over 30% of temporally restricted factoid questions without additions or modifications to the proposed approach.

We have begun to work in two directions: first the inclusion of additional features that allow us to respond questions whose answer is not necessarily expressed as a reformulation of the question into the target documents. Currently our work in this direction is based on the syntactic analysis of the retrieved passages, and in the inclusion of external knowledge. The second direction of research is the automatic selection of features *online* in order to get the best performance of the system for a given question.

**Acknowledgements.** This work was done under partial support of CONACYT (Project Grants U39957-Y and 43990), SNI-Mexico, and the Human Language Technologies Laboratory at INAOE. We also want to thank the CLEF as well as EFE agency for the resources provided.

## References

1. Carreras, X. and Padró, L. *A Flexible Distributed Architecture for Natural Language Analyzers.* In Proceedings of the LREC'02, Las Palmas de Gran Canaria, Spain, 2002.
2. Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., Rosso P. *A Passage Retrieval System for Multilingual Question Answering.* 8th International Conference on Text, Speech and Dialog, TSD 2005. Lecture Notes in Artificial Intelligence, vol. 3658, 2005.
3. Magnini B., Vallin A., Ayache C., Erbach G., Peñas A., Rijke M., Rocha P., Simov K., Sutcliffe R., *Overview of the CLEF 2004 Multilingual Question Answering Track*. In Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England, ISTI-CNR, Italy 2004.
4. Pérez-Coutiño M., Solorio T., Montes-y-Gómez M., López-López A. and Villaseñor-Pineda L., *Toward a Document Model for Question Answering Systems*. In Advances in Web Intelligence. Lecture Notes in Artificial Intelligence, vol. 3034, Springer-Verlag 2004.
5. Pérez-Coutiño M., Solorio T., Montes-y-Gómez M., López-López M. and Villaseñor-Pineda L., *Question Answering for Spanish Supported by Lexical Context Annotation*, In Multilingual Information Access for Text, Speech and Images, Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Peters C, et al. (Eds.), Lecture Notes in Computer Science, vol. 3491, Springer 2005.
6. Ravichandran D. and Hovy E. *Learning Surface Text Patterns for a Question Answering System*. In ACL Conference, 2002.
7. Saggion, H. *Identifying Definitions in Text Collections for Question Answering*. LREC 2004.