# Experiments for tuning the values 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 by the Language Technologies Laboratory at INAOE for Spanish monolingual QA evaluation task at CLEF 2005. Our approach is centered in the use of lexical features in order to identify possible answers to factual questions. Such method is supported by an alternative one based on pattern recognition in order to identify candidate answers to definition questions. The methods applied at different stages of the system and prototype architecture for question answering are described. The paper shows and discusses the results achieved with this approach.

**Keywords:** Question Answering for Spanish, Lexical Context, Natural Language Processing.

## 1 Introduction

Current information requirements claim for efficient mechanisms capable of interact with users in a more natural way. Question answering systems has been proposed as a feasible option for the creation of such mechanisms and the research in this field presents a continue growing both in interest as well as in complexity [3]. This paper presents the prototype developed at the Language Technologies laboratory of INAOE for the Spanish monolingual QA evaluation task at CLEF 2005. The experiments performed this year by our group continue with its last year work [5] in the following points, a) the approach is centered in the analysis of the lexical context related to each named entity selected as candidate answer; b) the information used to discriminate candidate and final answers relies on a shallow NLP processing (POS and named entities tagging) and statistical factors. On the other hand, there are some important modifications in the prototype architecture that allow the system to have a better performance (recall) at the initial stages. At the same time there have been some simplifications in the general architecture. For instance, we have made a shallow question classification process; and the answer discrimination process relies only on the information located in the target documents. Thus discarding the internet searching and extracting modules.

The paper is focused in the discussion of the processes involved in factual question answering. Nevertheless it is presented a section with the description of the methods used for answer definition questions. The rest of this paper is organized as follows; section two describes the architecture of the prototype; section three to section six details the internal processes of the system; section seven discusses the results achieved by the system; and finally section eight exposes our conclusions and discusses further work.

## 2 Prototype Architecture

The system is based on our last year methodology [5] but with some significant modifications in the prototype. Figure 1 shows the main blocks of the system. It could be observed that the treatment of factoid and definition questions occurs independently. Factoid questions resolution relies on a hybrid system involving the following stages: *question processing*, which includes the extraction of named entities and lexical context in the question, as well as question classification to define the semantic class of the answer expected to respond to a given question; *documents 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 the question processing); *searching*, where a set of candidate answers is obtained from the *modeled* passages retrieved by the PRS; and finally *answer extraction*, where candidate answers are 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 methodology supported by a couple of lexical patterns that allow finding and selecting the set of possible answers. Next sections describe each of these stages.

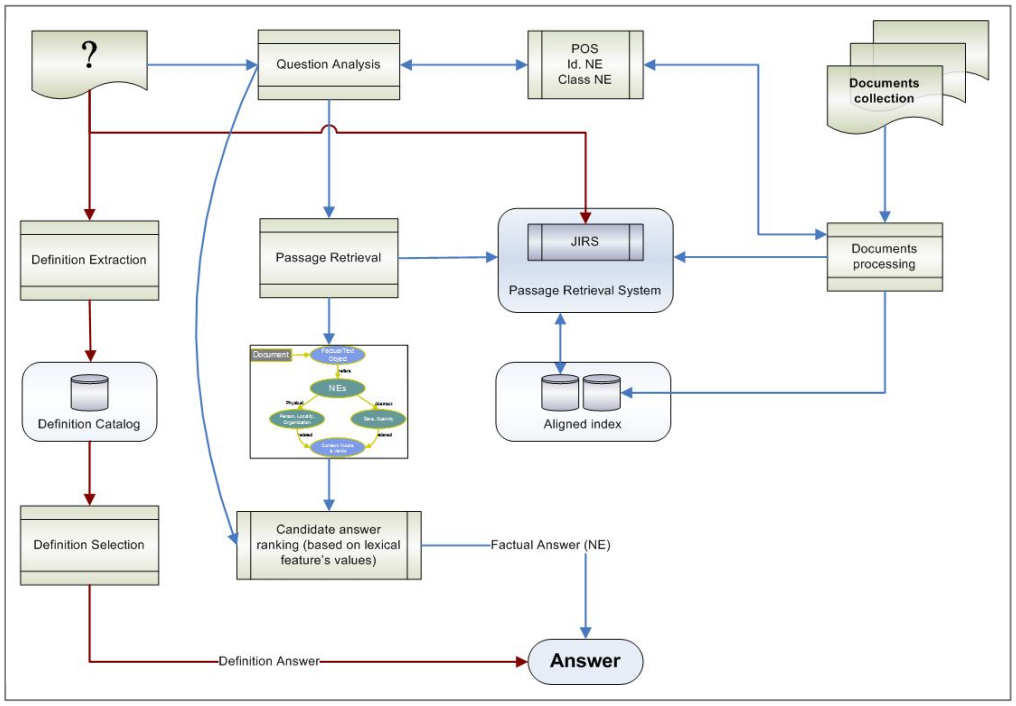


**Figure 1.** Block diagram of the system. Factoid and definition questions are treated independently. Factual questions require the following stages: question processing, documents processing, searching and answer selection. Definition questions use a pattern approach for definition extraction and definition selection process.

## 3 Question Processing

QA systems traditionally perform a question processing stage in order to know in advance the type (semantic class) of the answer expected by a given question and thus, reduce the searching space to only those information fragments related to the semantic class found. Our prototype implements this stage following a direct approach involving the next 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. Question's named entities are identified and classified using MACO.

The first step is responsible of identify the semantic class of the expected answer. In the experiments performed with the training data set, we found 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 use more than five classes, for instance person, organization, location, date, quantity and other. Steps 2 and 3 produce information that is used later in searching stage to match questions and candidate answer context, contributing to the weighted schema.

## 4 Documents Processing

This year we experiment with a *hybrid*[1] approach for document processing that has allowed simplifying greatly this stage. The processing of target documents is composed 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), which create the index for the searching process. The index gathered by JIRS and the tagged collection are aligned phrase by phrase for each document in the collection. This way, the system could retrieve later the relevant passages for a given question with JIRS, and then use their tagged form for the answer extraction process.

[1] The qualification of *hybrid* to this approach means that the system combines shallow NLP information (POS and named entity tagging) and statistical, language independent information encapsulated into JIRS.

## 5 Searching

Due to the document processing stage, searching stage is also performed in two steps. As we mention, 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 previous processing.

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

Assume that the user question is "*Who is the president of Mexico?*" and that two passages were obtained with the following texts: "*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 without the question word *Who*) the following sets are gathered:

**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".

Next, 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 one 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 give us a lower value than for $p_1$ because the second passage is very different with respect to the original question, although it contains all the relevant terms of the question.

Previous evaluation of JIRS also demonstrates that it is possible to achieve coverage of over 60% for the first 20 passages. That is, the possible answer to a given question is found between the first 20 passages retrieved by JIRS for over 60% of the training set. We refer the reader to [2] in order to get a complete discussion of the similarity metrics used by JIRS and its evaluation.

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 disadvantage of a couple of extra hours for such processing.

Tagged passages are represented in the same way that in [4] where each retrieved passage is modeled by the system as a factual text object whose content refers to several named entities even when it is focused on a central topic. As mentioned, named entities could be one of these: persons, organizations, locations, dates, quantities and miscellaneous[2]. 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 question's representation with the best set of candidates gathered from passages.

## 6 Answer Extraction

Answer extraction is performed according to the type of question, factual or definition. Next subsections detail the processes involved in answering each one.

### 6.1 Answering Factoid Questions

The system makes no difference between factual and temporal restricted factual 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 coefficient gather by 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, allowing to the final answer selection to rely on no more information

[2] The semantic classes used rely on the capability of the named entity classifier used in our experiments.

that just 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.

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

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

Where $\omega_A$ is the assigned weight for a candidate answer; $t_q$ is 1 if the semantic class of the candidate answer is the same that the question's one and 0 in other case; $n$ is a normalization factor based on the number of activated features, $NE_q$ is the set of named entities in the question and $NE_A$ is the set of named entities in the context of the candidate answer; $C_q$ is the question's context and $C_A$ is the candidate answer's context; $F_A(P_i)$ 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 by JIRS; and $1 - \frac{P_i}{k-1}$ is an inverse relation for the passage ranking.

**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 if the passages without question's named entities will be 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 sort order, taking as answer the one with the greatest weight. Section 7 describes some experiments performed with training data set and the results achieved with both, training and test sets.

### 6.2 Answering Definitions

Our system uses an alternative method to answer definition questions. This method makes use of some regularities of language and some stylistic conventions of news letters 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 to 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 letters 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 other one consists of a list of referent-description couples.

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

$$w_1 \langle meaning \rangle \; ( \; \langle acronym \rangle \; ) \qquad \text{(i)}$$

In this pattern, $w_1$ is a lowercase non stopword, <*meaning*> is a sequence of words starting with an uppercase letter (that can also include some stopwords), 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 particular this pair was catch from the following paragraph:

> "*El Partido Auténtico de la Revolución Mexicana (PARM) nombró hoy, sábado, a Álvaro Pérez Treviño candidato presidencial de ese organismo para las elecciones federales del 21 de agosto de 1994*".

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.

$$w_1 \; w_2 \; \langle description \rangle \; , \; \langle referent \rangle \; , \qquad \text{(ii)}$$

Where $w_1$ may represent any word, except for 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 being in the stopwords list.

Applying this extraction pattern over the below paragraph we could find the pair [*Alain Lombard - El director de la Orquesta Nacional de Burdeos*].

> "*El director de la Orquesta Nacional de Burdeos, Alain Lombard, ha sido despedido por el Ayuntamiento de esta ciudad, que preside Alain Juppé, tras un informe que denuncia malos funcionamientos y gastos excesivos*".

### 6.2.2 Definition Selection

The main quality 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. For instance, when using the extraction pattern (i) to analyze the following news we obtain the incorrect definition pair [Ernie - AFS]. In this case the resultant pair does not express an acronym-meaning relation; instead it indicates a person-nationality association.

> *Ernie Els (AFS) se mantiene en cabeza de la lista de ganancias de la "Orden de Mérito" de golf, con más de 17 millones de pesetas, mientras que el primer español es Miguel Angel Martín, situado en el puesto decimoséptimo, con 4.696.020.*

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 probable answer for a given definition question. The proposed approach is supported on the idea that, on the one hand, the correct information is more abundant than the incorrect one, and on the other hand, 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. Assume that 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 probable answer. This decision is based on the fact that this answer is the more frequent description related to Félix Ormazabal in the catalog.

*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 definitions in all our runs for monolingual QA, including those described in "INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering", the discussion on these results has been documented in that paper. Thus the rest of this document is intended to discus the results on factual question answering.

### 7.1 Training Experiments

As we mention earlier, the approach used in our system is similar to the one used last year [5], an analysis of such system show us that it was necessary to experiment with different values for the parameters involved in the answer extraction stage (see table 1). For instance, last year the system relied in a document model considering only four elements (just nouns and/or verbs) at left and right for the named entities context. This year we performed several experiments using context lengths from four elements to the whole passage retrieved, we also experiment with different elements: 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 factual question answering.

**Table 2.** Configurations of some experiments performed with training set.
First column corresponds to the feature list showed 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 | Pasaje | 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 reached with the "Exp. 7" which combines the following feature values, first the system classify 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 makes the proper representation for each named entity found in it. At this step, the context is formed by 8 elements at left and/or right of the named entity, and considers verbs, common names, named entities and adjectives. The extraction stage filters those candidates answers whose context does not contain at least one of the question's 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 methodology. 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 participate in the evaluation with two runs, both were gathered 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 factual questions) were over 10% better than last year, we believe that the approach described is near to its limits of accuracy. A shallow analysis of the results shows that the proposed system is suited for questions with some stylistic characteristics whose answer is commonly found in the near context of some reformulation of the question into the passages. While for others, more elaborated factual questions is unable to identify the right answer. That is the case of questions whose expected answer is an object or some abstract entity which can not be identified *a priori* by a shallow NLP 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 yield to a wrong selection of an answer.

A detailed analysis of these results will help us to take the next direction in our research.

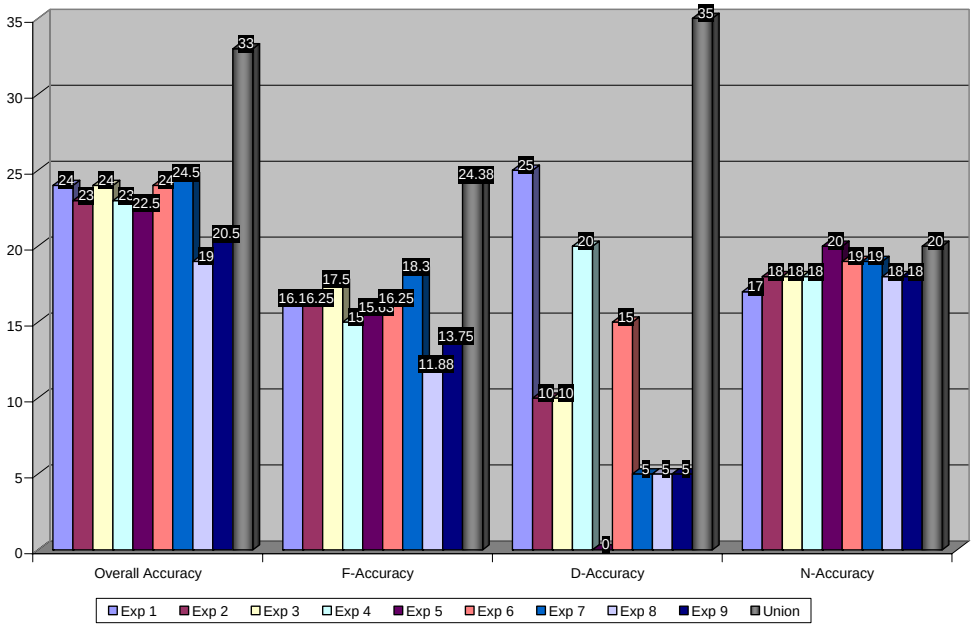


**Figure 2.** Results achieved with training set, applying the configurations showed in table 2.

We have begun some experiments in order to get the right configuration for each question online, that is, to select automatically the appropriate configuration for a given question based on question's attributes. Another direction in our research is to include more features that allow us to perform a better selection and discrimination of candidate answers, more over, that allow to consider objects and abstract entities that are currently excluded by the methodology.

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

After a shallow analysis of these results we have begun to work in two directions: first the inclusion of other 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 study of a syntactic ana-

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 given a 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 of INAOE. We also like to thanks to 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*, to appear in the 8th International Conference on Text, Speech and Dialog, TSD, Springer LNAI, 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. LNAI3034 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", to appear in proceedings of the 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Peters C, et al. (Eds.), September 2004, Bath, England, Springer-Verlag 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.