Mexican Experience in Spanish Question Answering
Experiencia Mexicana en la Búsqueda de Respuestas en Español

Manuel Montes y Gómez, Luis Villaseñor Pineda and Aurelio López López
Laboratorio de Tecnologías del Lenguaje, Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).
Luis Enrique Erro #1, Tonatzintla, Puebla, México.
mmontesg@inaoep.mx, villasen@inaoep.mx, allopez@inaoep.mx

Article received on April 14, 2008; accepted on June 20, 2008

Abstract.
Nowadays, due to the great advances in communication and storage media, there is more information available than ever before. This information can satisfy almost every information need; nevertheless, without the appropriate manage facilities, all of it is practically useless. This fact has motivated the emergence of several text processing applications that help in accessing large document collections. Currently, there are three main approaches for this purpose: information retrieval, information extraction, and question answering. Question answering (QA) systems aim to identify the exact answer to a question from a given document collection. This paper presents a survey of the Mexican experience in Spanish QA. In particular, it presents an overview of the participations of the Language Technologies Laboratory of INAOE (LabTL) in the Spanish QA evaluation task at CLEF, from 2004 to 2007. Through these participations, the LabTL has mainly explored two different approaches for QA: a language independent approach based on statistical methods, and a language dependent approach supported by sophisticated linguistic analyses of texts. It is important to point out that, due to these works, the LabTL has become one of the leading research groups in Spanish QA.
Keywords: Question Answering, Passage Retrieval, Answer Extraction, Machine Learning.

Resumen.
En la actualidad, debido a los grandes avances en los medios de comunicación y de almacenamiento, hay más información disponible como nunca antes se ha visto. Esta información puede satisfacer casi todas las necesidades de información, sin embargo, sin una adecuada gestión ésta es prácticamente inútil. Este hecho ha motivado la aparición de diferentes aplicaciones para el procesamiento de texto orientadas a facilitar el acceso a grandes colecciones de documentos. Hoy en día, existen tres enfoques principales para este propósito: la recuperación de información, la extracción de información, y los sistemas de búsqueda de respuestas. Los sistemas de búsqueda de respuestas (QA por sus siglas en inglés) tienen por objeto identificar la respuesta exacta a una pregunta dentro de una determinada colección de documentos. Este trabajo presenta un panorama general de la experiencia mexicana en QA en español. En particular, se presentan las participaciones del Laboratorio de Tecnologías del Lenguaje del INAOE (LabTL) en la tarea de QA en español dentro del foro de evaluación CLEF, desde 2004 a 2007. A través de estas participaciones, el LabTL ha explorado principalmente dos enfoques diferentes en QA: un enfoque independiente del lenguaje basado en métodos estadísticos, y un enfoque dependiente del lenguaje apoyado en un complejo análisis lingüístico del texto. Es importante señalar que, debido a estos trabajos, el LabTL se ha convertido en uno de los principales grupos de investigación de QA en español.
Palabras Claves: Búsqueda de Respuestas, Recuperación de Pasajes, Extracción de Respuestas, Aprendizaje Automático.

1 Introduction

Nowadays, due to the great advances in communication and storage media, there is more information available than ever before. This information can satisfy almost every information need; nevertheless, without the appropriate manage facilities, all of it is practically useless. This fact has motivated the emergence of several text processing applications that help in accessing large document collections. Currently, there are three main approaches for this purpose: information retrieval, information extraction, and question answering.

Information retrieval (IR) systems focus on searching relevant documents for general user queries. The output of this kind of systems is a ranked list of documents, where documents are ordered by their similarity to the query.

One of the main requirements for these systems is the ability to deal with huge amounts of textual information in very short times (in the order of milliseconds for a web search). This demand has imposed several restrictions on IR models, which are mainly based on statistical rather than on linguistic methods.

Information extraction (IE) systems consider the task of finding predefined pieces of information in a given –domain specific– document collection. The goal of an IE system is to find, extract and present such information in a structured format, typically as a database. In other words, these systems build a structured representation from unstructured textual information. These systems are assembled on demand for each specific task, depending on the kind of information to be extracted. As expected, these systems have to apply complex linguistic techniques in order to accurately detect and extract the required information.

IR and IE systems have made possible the processing of large volumes of textual information. However, they present serious problems for answering specific questions formulated by users. On one hand, IR systems can not tackle this task. In fact, once the user obtained a list of relevant documents for her question, she still has to review all documents in order to find the desired information. On the other hand, IE systems are more suitable to extract concrete information from document collections. Nevertheless, they require a precise definition of the kind of information to be extracted, and therefore, they do not allow the treatment of arbitrary questions.

All these inconveniences along with a growing need for improved information access mechanisms triggered the emergence of question answering (QA) systems. These systems aim at identifying the exact answer to a question from a given document collection. Evidently, QA requires efforts beyond information retrieval, since once a document about the question is found, the system has to go into the document content in order to locate the desired answer. This way, QA systems tend to use statistical IR methods to identify documents where the answer is likely to appear, as well as linguistic-based IE techniques to locate candidate answers.

The problem of answering questions has been recognized and partially tackled since the 70s for specific domains. However, with the advent of browsers working with billions of documents in Internet, the need has newly emerged, having led to approaches for open-domain QA. Some examples of such approaches are emergent question answering engines such as *answers.com*, *ask.com*, or additional services in traditional browsers, such as *Yahoo*.

Recent research in QA has been mainly fostered by the TREC[1] and CLEF[2] conferences. The first one focuses on English QA, whereas the second evaluates QA systems for most European languages except English. To date, both evaluation conferences have considered only a very restricted version of the general QA problem. They basically contemplate simple questions which assume a definite answer typified by a named entity or noun phrase, such as factoid questions (for instance, "*How old is Cher?*" or "*Where is the Taj Mahal?*") or definition questions ("*Who is Nelson Mandela?*" or "*What is the quinoa?*"), and exclude complex questions such as procedural or speculative ones.

This paper presents a survey of the Mexican experience in Spanish QA. In particular, it presents an overview of the participations of the Language Technologies Laboratory of INAOE (LabTL) in the Spanish QA evaluation task at CLEF, from 2004 to 2007. Through these participations, the LabTL has mainly explored two different approaches for QA: a language independent approach based on statistical methods, and a language dependent approach supported by sophisticated linguistic analyses of texts. It is important to point out that, due to these works, the LabTL has consolidated as one of the leading research groups in Spanish QA.

The rest of the paper is organized as follows. Section 2 introduces the QA task. Section 3 presents a brief survey of the Spanish QA evolution. Section 4 gives an overview of the Mexican experience in Spanish QA; basically, this section resumes the participation of the LabTL in the CLEF workshops from 2004 to 2007. Finally, Section 5 exposes our conclusions and describes some future work directions.

[1] Text Retrieval Conferences (trec.nist.gov/)

[2] Cross-Language Evaluation Forum (www.clef-campaign.org/)

2 Question Answering

As we previously mentioned, QA has become a promising research field whose aim is to provide more natural access to textual information than traditional document retrieval techniques. In essence, a QA system is a kind of information retrieval application that responds to natural language questions with concise and precise answers.

Question Types

Current work on QA has mainly focused on answering two basic types of questions: factoid and definition questions. Factoid questions usually require a simple fact (a date, a quantity, or the name of something) as answer, whereas definition questions ask for concrete descriptions of objects or persons commonly expressed by simple noun phrases. Table 1 shows some example questions from the CLEF evaluation tasks.

Table 1. Example questions from the QA@CLEF task

Question	Category	Answer	Answer Type
What year was Thomas Mann awarded the Nobel Prize?	Factoid	1929	Date
How long is the Freixo Bridge?	Factoid	700 meters	Quantity
Where is the Auschwitz death camp located?	Factoid	Poland	Location
Who was the first president of the United States?	Factoid	George Washington	Person
Who is Yves Saint Laurent?	Definition	fashion designer	Definition
What is the Mossad?	Definition	Israeli Secret Service	Definition
What is the quinoa?	Definition	gluten-free grain cultivated by the Incas	Definition

Architecture of QA systems

From the QA systems developed so far, it is possible to identify a prototypical architecture that consists of three main modules: question analysis, passage retrieval and answer extraction.

The role of the question analysis module is to identify the expected answer type, i.e., the semantic category to which the answer belongs. For instance, when asking "*Where did a meteorite fall in 1908?*", the expected answer type is "Location". In addition, this module is also intended to identify the question words that can be used to query the document collection in order to find relevant passages.

The purpose of the second module, the passage retrieval, is to retrieve from the given document collection the set of passages –commonly considered as text paragraphs– more likely to contain the answer.

Finally, the answer extraction module uses the subset of passages obtained by the previous module and the expected answer type identified in the first one, to locate the exact text string representing the answer to a given question.

Section 3 enumerates some techniques applied in Spanish QA for these three subtasks.

Evaluation of QA systems

The evaluation of QA systems has been mainly carried out by the TREC and CLEF conferences; being the latter responsible of evaluating the development in Spanish QA. The Spanish QA evaluation task of CLEF (known as Spanish QA@CLEF) has been undertaken since 2003, and consists of answering 200 questions, mainly factoid and definition questions, from a document collection of 454,045 news reports from the Spanish EFE agency.

The goal of the QA@CLEF task is to define a common framework for the evaluation of QA systems, and therefore, to perform an objective comparative evaluation of them. In order to do that, all participating systems use the same search collection, the same set of test questions, and they are evaluated using the same criteria. In a few words, the evaluation of QA systems at the Spanish QA@CLEF consists of the following steps:

- Each participating group delivers to the task organizers the set of answers to the questions, produced by its system.

- The supplied answers are judged by human assessors, who assign to each response a unique label, either right or wrong. An answer is judged to be right when it consists of nothing more than the exact, minimal answer and when the returned passage supports the response. In any other case, the response is considered to be wrong.
- The organizers compute the system accuracies as the proportion of questions which were correctly answered.

3 Spanish Question Answering

This section presents a brief survey of Spanish QA. It mainly shows the results from the Spanish QA@CLEF track from 2003 to 2007 [23, 24, 41, 25, 19], and describes the main ideas developed for the task through all these years.

As we previously mentioned, the QA@CLEF task has fostered the research in Spanish QA. This fact is attested in Table 2, which enumerates the task's participants per year. This table also evidences that the Language Technologies Laboratory of INAOE (LabTL) has been the only participant from Latin America.

Table 2. Participants of the Spanish QA@CLEF task

Year	Participants
2003	University of Alicante [42]
2004	University of Alicante [43], University Carlos III [12], University of La Coruña [26], Polytechnic University of Cataluña [1], *INAOE* [28].
2005	University of Alicante [34, 39], University Carlos III [13], Polytechnic University of Valencia [21], Polytechnic University of Cataluña [17], *INAOE* [30, 27].
2006	University of Alicante [16, 40], University Carlos III [14], Polytechnic University of Valencia [7], Polytechnic University of Cataluña, Priberam Informatica [9], University of Jaén [18], *INAOE* [31, 22].
2007	University Carlos III [15], Polytechnic University of Valencia [8], Polytechnic University of Cataluña, Priberam Informatica [3], *INAOE* [37].

In the same way, Table 3 shows the evolution of the Spanish QA research in the framework of the CLEF. It includes the best results from 2003 to 2007. It is evident that there has been a significant advance in the last two years, where overall accuracy was about 50% and accuracy on definition questions has almost reached 90%. Regarding these results, it is important to emphasize the role of the LabTL: it obtained the best overall accuracy in the 2005 evaluation exercise and has achieved the best results for definition questions in the last three years.

Table 3. Evolution of best performing systems at Spanish QA@CLEF

Year	Best overall accuracies	Best accuracies in factoid questions	Best accuracies in definition questions
2003	24.5	24.5	-
2004	32.5	31.0	70.0
2005	42.0	29.6	80.0
2006	52.5	55.5	83.3
2007	44.5	47.8	87.5

Next we present the main approaches developed for Spanish QA. Mainly, we enumerate most common techniques used in each one of the three QA subtasks: question classification, passage retrieval and answer extraction.

In question classification, it is possible to distinguish two major approaches. One of them is based on the application of lexical-syntactic patterns [42, 17, 27, 30, 16, 14, 3], and the other considers the use of machine learning techniques [39, 21, 40, 7, 18]. In both cases, one important difference among systems is the number of considered answer types. Some groups only consider 3 basic types (date, quantity and name) [27, 30, 31, 22, 37]; while the rest work with several types and even some subtypes. For instance, [16] use 200 answer types and [7, 8] consider a three-level type hierarchy.

Passage retrieval has been mainly done using traditional IR engines, such as Lucene[3], Lemur[4] and Xapian[5]. Nevertheless, there were also some efforts related to the development of retrieval systems specially suited for QA, that is the case of IR-n [42, 43, 39, 40, 16, 18] and JIRS [21, 27, 30, 7, 31, 22, 8] passage retrieval systems.

Additionally, some groups have explored some advanced ideas for passage ranking; for instance, [40] used the LSI technique for a semantic re-ranking, and [18] applied information fusion strategies to combine into one single ranked list the output of different retrieval engines. Also, some other groups have achieved passage retrieval not only from the EFE collections but also from the Web. In particular, [42, 43, 39] explored the retrieval of information from the English and Spanish Web.

In the case of answer extraction, one can distinguish two different approaches. The first is exclusively based on statistical techniques [28, 39, 27, 30, 21, 7, 22, 14, 18, 37, 8]. It determines the question's answer mainly based on: (i) the overlap between question words and the answer's context, and (ii) the frequency of occurrence of the answer in the given set of passages. On the other hand, the second approach not only considers these statistical criteria but also takes into account some syntactic and semantic restrictions on the answer's context [1, 17, 34, 16, 9, 31, 3]. Most systems perform a "heuristic" combination of these criteria, except by [22] that applied a machine learning approach.

Finally, it is important to mention that only three groups (namely, Polytechnic University of Cataluña, Priberam and INAOE) designed methods specially suited for the treatment of definition questions. In particular, the first two groups [17, 9, 3] used manually defined lexical-syntactic patterns, whereas the LabTL [22, 31] applied automatically discovered lexical patterns.

4 LabTL participations at QA@CLEF

This section describes the participations of the Language Technologies Laboratory of INAOE (LAbTL) in the Spanish QA@CLEF task from 2004 to 2007. In all these participations, our systems were based in the general architecture showed in Figure 1, which considers two major modules. The first focuses on answering factual questions. It considers the tasks of passage indexing, where documents are preprocessed and a structured representation of the collection is built; passage retrieval, where the passages with more probability to contain the answer are found from the index; and answer extraction, where candidate answers are ranked and the final answer recommendation of the system is produced. The second module concentrates on answering definition questions. It includes the tasks of definition extraction, where all possible pairs of name-description[6] are located and indexed; and definition selection, where relevant data pairs are identified and the final answer of the system is generated.

[3] lucene.apache.org/

[4] www.lemurproject.org/

[5] www.xapian.org/

[6] QA@CLEF tasks have considered three kinds of definition questions; some asking for the definition –general description– of a person, some others asking for the description of an acronym and others requesting a definition of an object (refer to table 1).

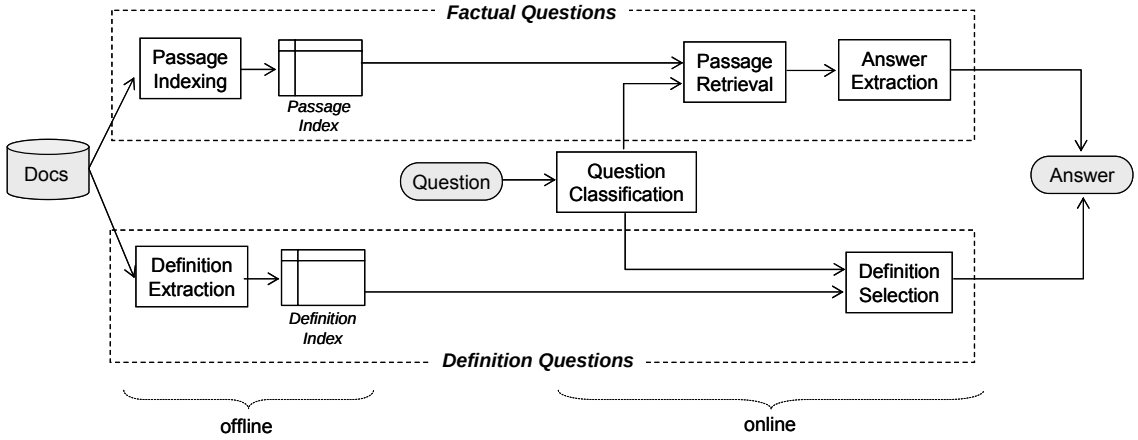

Fig. 1. General architecture of our QA systems

The following subsections present the systems for each of our participations at the Spanish QA@CLEF. They describe the main details of each system, as well as the evaluation conditions and the achieved results.

4.1 CLEF 2004

4.1.1 System overview

For our first participation at QA@CLEF, we adopted two approaches that were successfully evaluated in English at previous TREC campaigns. On one hand, we applied a predictive annotation approach [32] and, on the other hand, we used the Web as additional information source for answer validation [6]. Our purpose was to investigate the usefulness of these techniques in Spanish QA.

In particular, we applied these techniques for answering factoid questions. Based on the idea of predictive annotation, we considered all named entities from the target collection as possible answers. Then, given a question, we compared it against all entity contexts and then selected the entities with the highest similarity as candidate answers. Subsequently, these candidate answers were compared against a second set of answers gathered from the Web –using the method proposed by Del-Castillo et al. [10], and the final responses were selected based on a set of relevance measures, which encompass all the information collected in the searching process.

On the other hand, for answering definition questions our system implemented a pattern matching algorithm as those described in [33, 35]. This simple method used some regularities of language and some stylistic conventions of news reports to capture the possible answer for a given definition question.

The following section details the main components of the system of this year. A complete description can be found in [28].

4.1.2 System components

Question classification

This module allowed distinguishing among six different answer types: person, organization, place, date, quantity and miscellaneous. In order to do that, it applied a straightforward supervised approach, where the attributes for the learning task were the question words (including the interrogative particle), along with additional information obtained from the Web. In particular, it was implemented by a Support Vector Machine and was trained with the questions from the CLEF-2003 evaluation.

The main innovation of this module was the introduction of information from the Web. The procedure for gathering this information considered the following steps. First, we used a set of heuristics to detect the main noun (focus) of the given question. Then, we searched the Web using some queries that combine the question focus and the five (non-ambiguous) possible answer types. For instance, for the question *Who is the president of the french republic?*, president was defined as the question focus, and the following five queries were built: "*president is a*

person", "*president is a place*", "*president is a date*", "*president is a measure*", and "*president is an organization*". Finally, using the web results, we computed five additional features that express the relation of the question focus with each of the expected answer types.

The idea behind this procedure was to obtain some evidence about the usage of the question focus in the Web. In some cases, this evidence is very valuable; for instance, in our example, the first query was the most abundant, indicating that *person* is the most reliable answer type for the question at hand. Unfortunately, this evidence is not always as obvious as in this case. In order to handle this situation, we decided to compute some additional features that express the relation of the question focus with each of the expected answer types, and then we employed a learning algorithm to take advantage of this information.

Passage retrieval

As we previously mentioned, in a predictive annotation approach the indexing units are the named entities. Following this idea, each document of the collection was described by a set of named entities and their lexical contexts. Particularly, our system prototype considered six types of named entities (persons, organizations, locations, dates, quantities and miscellaneous) and represented their contexts by their surrounding nouns and verbs. Table 4 shows an example of a named entity and its context. For details about the document model and its automatic construction, we refer the reader to [29].

Table 4. Context associated to the named entity "CFC"
Underlined verbs and common nouns indicated the lexical contexts of the named entity

<DOCNO>EFE19941219-11009</DOCNO> … son usados en los productos anticongelantes, de insuflación y como refrigerantes, que tienen al cloro como un ingrediente común. "Los **CFC** son los responsables del agujero de la capa de ozono",…

According to the proposed document model, the indexing step considered the storage of the extracted named entities and their related information (its semantic class, associated contexts, and the documents where they appeared) in a relational database, whereas the retrieval step only consisted of extracting from this database all named entities having the expected answer type and a context overlap with the words of the question.

Answer extraction

The answer extraction procedure included two basic tasks. First, answers were filtered using the nformation obtained from the Web. The idea was to eliminate candidate answers that could not be extracted from the Web (using the approach described in [10]). Second, candidate answers were analyzed in order to select from them the answer most likely to satisfy the question. For this task, we considered five criteria: 1) the number of times the candidate answer was labeled with the expected answer type C_r; 2) the number of times the candidate answer was labeled with a different entity type C_w; 3) the cardinality of the intersection between the contexts of the answer and the question, excluding named entities I_t; 4) the number of context entities that matched the entities in the question I_{en}, and 5) the frequency of the candidate answer in the whole collection of documents F_{a_i}. In particular, we computed the weight of each candidate answer a_i based on formula (1), where coefficients α y β were used to compensate the errors from the named entity recognition process. At the end, we selected the candidate answer with the highest weight as the final answer.

$$w(a_i) = \alpha C_r + \beta C_w + I_t + I_{en} + F_{a_i} \quad (1)$$

Definition extraction and selection

The definition extraction module used only two basic typographic extraction patterns to detect the text segments that contain the description or meaning of a term. These patterns mainly focused on the extraction of information related to organizations and persons (such as their acronyms and their positions). In particular, they allowed to gather two

definition catalogs: a list of acronym-description pairs, and a list of person-description pairs. In order to extract the acronym-description pairs, we used an extraction pattern based on the use of parentheses.

$$w_1 \text{ <description> (<acronym>)}$$

In this pattern, w_1 is a lowercase word, <*description*> 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 caught from the following passage:

> *"...el Partido Auténtico de la Revolución Mexicana (PARM) nombró hoy, sábado, a Álvaro Pérez Treviño candidato presidencial de ese organismo ..."*

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

$$W_1\ w_2 \text{ <description> , <person> ,}$$

Where w_1 represents any word, except by a preposition, w_2 is a determiner, <*description*> is a free sequence of words, and <*person*> indicates a sequence of words starting with an uppercase letter or being in the stopword list. Applying this extraction pattern over the passage below, we could find the pair [*Andrew Lack - presidente del departamento de noticias de la cadena NBC*].

> *"... tanto es así, que el presidente del departamento de noticias de la cadena NBC, Andrew Lack, confesó en una entrevista ...".*

Once collected both catalogs, they were used by the definition selection module to extract the answer for a given question. That is, given a question, this module firstly identified the question target term, and then it extracted the first definition associated to this term from the catalogs.

4.1.3 Evaluation results

In summary, our QA system evaluated at the Spanish QA@CLEF 2004 was based on a predictive annotation approach, where documents where indexed using the lexical context of their named entities. It also included a method for discriminating candidate answers. In particular, it considered merging evidence from two different sources: a ranked list of local candidate answers, and a set of candidate answers gathered from the Web.

Table 5 shows the evaluation results of our system. An analysis of its errors revealed some obvious disadvantages of this approach. On one hand, the predictive annotation greatly depends on the precisions of both the named entity recognition process, and the question classification module. The mistakes in these early stages produced an error cascade that affects the final answer accuracy. As a fact, from the 200 test questions, our question classification method could only assign the correct answer type to 157 questions, which represent a precision of 78.5%. On the other hand, we noticed that the used contexts were not relevant for several entities. In this line, it is necessary to perform additional experiments in order to determine the most appropriate number and kind of elements for representing them. Specifically, it is important to evaluate the impact caused by using larger contexts, as well as considering other kind of elements, such as adjectives and adverbs.

Table 5. Results of our system at QA@CLEF 2004

	No. questions	Right	Accuracy
Definition	20	10	50.0%
Factoid	180	35	19.4%
TOTAL	200	45	22.5%

4.2 CLEF 2005

Based on the lessons learned from the previous evaluation exercise, this year we explored two different ideas for answering factoid questions: one of them extended our 2004 work and the other started our experimentation on a full data-driven approach. In consequence, the LabTL evaluated two different QA systems at Spanish QA@CLEF 2005. Both systems used the same passage retrieval method as well as the same module for answering definition questions. The following subsections describe the main elements from both systems.

4.2.1 System 1: A named entity centered approach

This system continued our previous year work in the following points: (i) the approach was centered in the analysis of lexical contexts related| to each named entity, and (ii) the information used to discriminate candidate answers relied on a shallow NLP processing (POS and named entities tagging) and some statistical factors. Nevertheless, because of the enormous dependence of the predictive annotation approach upon the named entity recognition process, for this year we decided to include a passage retrieval system. That is, given a question, we first retrieved a set of relevant passages; then, we –exclusively– analyzed the retrieved passages in order to identify their named entities (i.e., candidate answers); and finally, we compared the question against the lexical context of each candidate answer and selected as final answer that with the highest similarity.

In addition, with the aim of reducing the errors caused by wrong named entity and question classifications, we also decided to generalize the types of named entities. Mainly, we only considered three basic named entity types: date, quantity and proper name.

Finally, it is important to mention that we also eliminated the dependence of our method to external information sources. In particular, we discarded the use of web information for the answer discrimination process.

A complete description of the system can be found in [30].

4.2.1.1 System components

Question Processing

Previous year experiments exposed us the problems for building a robust method for question classification, and also revealed the great impact of its errors over the final answer accuracy. In order to improve the performance of this module we made the following simplifications: (i) we only considered three basic answer types (date, quantity and name); and (ii) we implemented it by means of a set of lexical patterns instead of by a machine learning approach.

Passage retrieval

The passage retrieval (PR) method used this year (in both systems) was specially suited for the QA task [20]. It allowed retrieving the passages with the highest probability to contain the answer, instead of simply producing the passages sharing a subset of words with the question. This method works as follows:

Given a user question, the PR method finds the passages with the relevant terms (non-stop words) using a classical information retrieval technique based on the vector space model. Then, it measures the similarity between the *n*-gram sets of the passages and the user question in order to obtain new weights for the passages. The weight of a passage is related to the largest *n*-gram structure of the question that can be found in such passage. The larger the *n*-gram structure, the higher the weight of the passage. Finally, it returns the passages with the new weights.

The similarity between a passage d and a question q is defined by (2).

$$sim(d,q) = \frac{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h\left(x(j), D_j\right)}{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h\left(x(j), Q_j\right)} \tag{2}$$

Where $sim(d, q)$ is a function which measures the similarity of the set of *n*-grams of the question q with the set of *n*-grams of the passage d. Q_j is the set of *j*-grams that are generated from the question q and D_j is the set of *j*-grams of the passage d. That is, Q_1 will contain the question unigrams whereas D_1 will contain the passage unigrams, Q_2

and D_2 will contain the question and passage bigrams respectively, and so on up to Q_n and D_n. In both cases, n is the number of question terms.

The result of (2) is 1 if the longest n-gram of the question is found in the set of passage n-grams. The function $h(x(j), D_j)$ measures the relevance of the j-gram $x(j)$ with respect to the set of passage j-grams, while the function $h(x(j), Q_j)$ is a factor of normalization[7]. The function h assigns a weight to every question n-gram as defined in (3).

$$h\left(x(j), D_j\right) = \begin{cases} \sum_{k=1}^{n} w_{\hat{x}_k(1)} & \text{if } x(j) \in D_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Where the notation $\hat{x}_k(1)$ indicates the k-th unigram included in the j-gram x, and $w_{\hat{x}_k(1)}$ specifies the associated weight to this unigram. This weight gives an incentive to the terms –unigrams– that appear rarely in the document collection. Moreover, this weight should also discriminate the relevant terms against those (e.g. stop words) which often occur in the document collection. The weight of a unigram is calculated by (4):

$$w_{\hat{x}_k(1)} = 1 - \frac{\log\left(n_{\hat{x}_k(1)}\right)}{1 + \log\left(N\right)} \tag{4}$$

Where $n_{\hat{x}_k(1)}$ is the number of passages containing the unigram $\hat{x}_k(1)$, and N is the total number of passages in the collection. We assume that the stop words occur in every passage (i.e., n takes the value of N). For instance, if the term appears once in the passage collection, its weight will be 1 (the maximum weight), whereas if the term is a stop word, then its weight will be the lowest.

Answer extraction

In order to achieve the answer extraction, the retrieved passages were represented in the same way that in our previous year work, that is, each passage was modeled by a set of named entities and their contexts (i. e., the nouns and verbs around the named entity). However, this year we used an improved formula for evaluating the relevance of contexts. We computed the weight for each candidate answer (named entity) based on coefficient (5). In this case, the candidate answer a_i having the highest weight was selected as the final answer.

$$w_{lex}(a_i) = \frac{t_q}{4} \times \left(\frac{\left|NE_q \cap NE_{a_i}\right|}{\left|NE_q\right|} + \frac{\left|C_q \cap C_{a_i}\right|}{\left|C_q\right|} + \frac{F_{a_i}(p_i)}{F_{a_i}(P)} + \left(1 - \frac{i}{k-1}\right) \right) \tag{5}$$

Where $w_{lex}(a_i)$ is the assigned weight for the candidate answer a_i; t_q is 1 if the semantic class of the candidate answer is the same that the that of the question, and 0 in other case; NE_q is the set of named entities in the question and NE_{a_i} is the set of named entities in the context of the candidate answer; C_q is the question context and C_{a_i} is the candidate answer context; $F_{a_i}(p_i)$ is the frequency of occurrence of the candidate answer a_i in the passage p_i; $F_{a_i}(P)$ is the frequency of occurrence of the candidate answer a_i in the whole set of retrieved passages; and $1 - \frac{i}{k-1}$ is an inverse relation for passage ranking. In this case, $i = 1..k$; k indicates the number of passages produced by JIRS.

Definition extraction and selection

For answering definition questions our system used an improved version of the pattern matching algorithm applied in 2004. In the same way, there were built two definition catalogs: person-description and acronym-description.

[7] We introduced the notation $x(n)$ for the sake of simplicity. In this case $x(n)$ indicates the n-gram x of size n.

As noticed in the previous year experiment, the main quality of the used extraction patterns was their generality. However, this generality caused the patterns to extract often non relevant information, i.e., information that does not indicate a relation acronym-description or person-description. Given that the catalogs contained a mixture of correct and incorrect relation pairs, it was necessary to do an additional process in order to select the most probable answer for a given definition question. The proposed approach was supported on the idea that, on one hand, the correct information is more abundant than the incorrect, and on the other hand, that the correct information is redundant. Thus, the process of definition selection was based on the following criterion: "the most frequent definition in the catalog has the highest probability to be the correct answer". In the case that two or more definitions have the same frequency, we decided by choosing the longest one, since this points to selecting the more specific definition, and therefore, the more pertinent answer.

The following example illustrates the process of answer selection. Given the question "*Who is Félix Ormazabal?*", and assuming that the definition catalog contains the records showed below. Then, our system would select the description "*diputado general de Alava*" as the most probable answer. This decision would be 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: diputación de este territorio:
Félix Ormazabal: presidente del PNV de Alava y candidato a diputado general:
Félix Ormazabal: diputado de Alava
Félix Ormazabal: través de su presidente en Alava
Félix Ormazabal: diputado general de Alava
Félix Ormazabal: diputado general de Alava

4.2.1.2 Evaluation results

Table 6 shows the evaluation results of this system. The achieved results improved our 2004 performance by over 10% on factoid questions and over 30% on definition questions. Moreover, this approach was able to answer over 30% of temporally-restricted factoid questions without additions or modifications to the proposed approach. In addition, this system obtained the best overall results for this year (refer Table 2 in section 3).

Table 6. Results of system 1 at QA@CLEF 2005

	No. questions	Right	Accuracy
Definition	50	40	80.0%
Factoid	150	44	29.3%
TOTAL	200	84	42.0%

As we previously mentioned, this first system was similar to the one used in 2004. An analysis of the preceding system showed us that it was necessary to experiment with different values for the parameters involved in the answer extraction stage. In 2004, the system relied on a document model considering only four elements (nouns and verbs) for describing the context of candidate answers. In contrast, this year our system obtained the best results using contexts of 8 elements, and taking into account not only common nouns and verbs, but also named entities and adjectives. In addition, this system computed the weight of each candidate answer by an improved formula.

Despite the relevant obtained results, we believed that the proposed approach was near to its limits of accuracy. An analysis of results showed that this system was suited for questions with some stylistic characteristics whose answer was found in the near context of some reformulation of the question into the passages. Whereas, for other more elaborated factual questions, the system was unable to identify the right answer. In this direction, our conclusion was to include syntactic information in the next year prototype.

4.2.2 System 2: Full data-driven QA

The majority of QA systems use a variety of linguistic resources to help in understanding the questions and the documents. The most common linguistic resources include: part-of-speech taggers, parsers, named entity extractors, dictionaries, and WordNet. Despite the promising results of these approaches, they have three main inconveniences: (i) the construction of such linguistic resources is a very complex task; and (ii) these resources are highly specific to a language; and (iii) the performance of these tools is still far from being perfect, especially for languages other than English.

The second system that we designed for the QA@CLEF 2005 was based on a full data-driven approach [6]. This system required minimum knowledge about the lexicon and the syntax of the specified language. Mainly, it was supported on the idea that the questions and their answers are commonly expressed using the same set of words, and therefore, it simply used a lexical pattern matching method to identify relevant document passages and to extract the candidate answers (for a complete description of the system refer to [27]).

Data-drive systems, such as the proposed, have the advantage to be easily adapted to several different languages, in particular to moderately inflected languages such as Spanish, English, Italian, and French. Unfortunately, this generality has its price. To obtain a good performance, the systems require a redundant target collection, that is, a collection in which the question answers occur repeatedly. On one hand, this redundancy increases the probability of finding a passage containing a simple lexical matching between the question and the answer. On the other hand, it enhances the answer extraction, since correct answers tend to be more frequent than incorrect responses.

4.2.2.1 System components

This second system only differs from the other in the way it performed the answer extraction for factoid questions, where it did not use any Spanish linguistic resource (e.g. a POS tagger or a lemmatizer). The rest of the modules, which did not rely on linguistic resources, were the same for both systems. Hence, this section exclusively details the answer extraction method, whereas the rest of modules are described in section 4.2.1.1.

Answer Extraction

This component aimed to determine the best answer for a given factoid question. In order to do that, it first selected a small set of candidate answers, and then, it picked the final unique answer taking into consideration the occurrences of the candidate answers in the retrieved passages. This method is based on the assumption that the retrieved passages are relevant, and therefore, that the correct answer will appear several times. However, this repetition could be inexact. For instance, if the answer is a date, one passage may contain the complete date whereas other only a fragment (e.g., only the year). In order to deal with this situation, our method compensated the frequency of fragmented answers by taking into consideration the frequency of the whole answer. The algorithm applied to calculate this compensated frequency for each candidate answer from the given set of relevant passages is described below. For more detail refer to [10].

1. Extract all the unigrams that satisfy a given typographic criteria. These criteria depend on the type of expected answer. For instance, if the expected answer is a named entity, then we select the unigrams starting with an uppercase letter. But if the expected answer is a quantity, then we select the unigrams expressing numbers.
2. Determine all the *n*-grams assembled from the selected unigrams. These *n*-grams can only contain the selected unigrams and some stop words.
3. Rank the *n*-grams based on their compensated frequency. The compensated frequency of the *n*-gram $x(n)$ is computed as follows:

$$F_{x(n)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n-i+1}\frac{f_{\hat{x}_j(i)}}{\sum_{\forall y \in G_i} f_{y(i)}} \tag{6}$$

where G_i indicates the set of *i*-grams, $y(i)$ represents the *i*-gram *y*, $\hat{x}_j(i)$ is the *j*-th *i*-gram included in $x(n)$, $f_{m(i)}$ specifies the frequency of occurrence of the *i*-gram *m*, and $F_{x(n)}$ indicates the compensated frequency of $x(n)$.

4. Select the top five *n*-grams as candidate answers.
5. Compute a ranking score for each candidate answer. This score is defined as the weight of the first retrieved passage (refer to formula 1) that contains the candidate answer.
6. Select as the final respond the candidate answer with the highest ranking score. In the case that two or more of the candidate answers have the same ranking score, then we select that with the highest compensated frequency.

4.2.2.2 Evaluation results

Table 7 shows the results of our full data-driven QA system. Notice that they are almost the same than those of the best 2005 QA system (see Table 2). These results were surprising given that this method did not use any linguistic resource. We believe this good performance was consequence of two conditions that were satisfied: a large document collection and a high redundancy of the answers.

Table 7. Results of system 2 at QA@CLEF 2005

	No. questions	Right	Accuracy
Definition	50	40	80.0%
Factoid	150	42	28.0%
TOTAL	200	82	41.0%

As we mentioned in previous sections, the definition catalogs were built using a quite naïve approach, and therefore they were incomplete and included several erroneous registers. Nonetheless, the achieved results (80%) indicated that they constituted a valuable information repository for answering definition questions.

It is also important to mention that we performed a series of experiments on Spanish, Italian and French in order to demonstrate the potential and portability of our system [27]. The results of such experiments confirmed the language independence of our system (it achieved the best result in Italian and the second best in French at QA@CLEF 2005). However, they also revealed that the method for answering factoid questions greatly depends on the redundancy of the answers in the target collection.

4.3 CLEF 2006

The participation of the LabTL in the Spanish QA@CLEF 2006 task once again involved two different systems. These systems extended the ideas explored in 2005 for answering factoid questions. In particular, the first system tested a new context representation based on the use of syntactic information. Its purpose was to enhance the answer extraction process. On the other hand, the second system applied a machine learning approach that facilitated the usage of a higher number of features for answer extraction. Both systems included the same component for answering definition questions. This last component improved our previous year work by applying extraction patterns that were automatically discovered, rather than manually defined. For question classification and passage retrieval, we used the same components than previous year (refer to section 4.2.1). The following sections describe the main modules of both systems.

4.3.1 System 1: Using syntactic features in QA

The analysis of previous year results indicated that different candidate answers could have equally relevant lexical contexts. This observation motivated us for searching richer context representations as well as robust weighting functions. In particular, we incorporated the following new elements to our system: (i) a syntactic parser based on dependency grammars for offline processing of the document collection, (ii) a shallow technique for weighting the question terms that have a syntactic dependency to one candidate answer within a relevant passage (aka term

density), and (iii) a linear combination for merging the term density of each candidate answer with the weights gathered in the previous steps to obtain its final weight.

This analysis also indicated that other important disadvantage of our previous year system was the manual definition of the extraction patterns, and consequently, the application of a very small number of such patterns for answering definition questions. Because of this disadvantage, we decided to generalize the treatment of definition questions by applying an automatic process for pattern discovery. The main purpose of this new process was to discover a greater number as well as a major diversity of answer extraction patterns.

The following subsection describes the details of the new modules of this system. A complete description of them can be found in [31].

4.3.1.1 System components

Answer Extraction

In order to improve the precision at the answer selection step, we experimented with the inclusion of a shallow approach based on the use of some syntactic features. There have been different approaches that use syntactic information for QA. Some of them have tried to find a direct similarity between the question and the answer structures. This kind of similarity is commonly supported by a set of syntactic patterns which are expected to match the answers. Some systems that applied this approach are described in [5, 34, 4]. Other proposals, as that presented by Tanev [36], have applied transformations to the syntactic tree of the question in order to approximate it to the trees of the relevant passages.

Despite the degree of effectiveness of those approaches, their principal drawback came up when the parsing was deficient. In order to cope with these deficiencies, we proposed a straightforward metric to weight candidate answers in accordance with its closeness to the question terms. Following we briefly describe the procedure to measure the syntactic density of a candidate answer in a passage.

Given a question $Q = \{q_1,...,q_n\}$ and a candidate answer a_i immersed in a passage $p = \{w_1,...,w_k\}$, we first constructed the dependency tree of the passage p, denoted as T_p. This tree could be represented by several sub-trees, since a passage can be formed by several sentences or could not be analyzed entirely. Then, we selected the sub-tree $t \in T_p$ that contains the candidate answer (we used the notation $<w, t>$ to indicate that the term w is part of the tree t). In this case, this sub-tree is defined as the context of the candidate answer a_i. Finally, we computed the syntactic density of a_i in t according to formula (7).

$$w_{syn}(a_i) = \frac{1}{n} \sum_{\forall q_i \in Q} f(q_i), \qquad (7)$$

$$f(q_i) = \begin{cases} 1 & \text{if } \langle q_i, t \rangle \\ 0 & \text{other case} \end{cases}$$

Due to the deficiencies on the construction of the dependency trees, we decided to combine all available information about a candidate answer a_i into one single measure. This way, the final weight of the i-th candidate answer, $w(a_i)$, was determined by combining its lexical and syntactic weights as follows:

$$w(a_i) = \alpha w_{lex}(a_i) + \beta w_{syn}(a_i) \qquad (8)$$

Where $w_{lex}(a_i)$ expresses the lexical weight of a_i (refer to Formula 5) and $w_{syn}(a_i)$ indicates its syntactic density. In this formula, coefficients α and β were experimentally defined to 1/3 and 2/3 respectively. In this way, the syntactic weight had a higher relevance.

Answering definition questions

Our previous method for answering definition questions was very precise; however, due to the application of a very small number of extraction patterns, it showed a low recall rate. That is, it could not respond to all definition

questions. In order to tackle this situation, in 2006 we applied an automatic process for discovering answer extraction patterns. In particular, our new method for answering definition questions[8] consisted of three main modules: a module for the discovery of extraction patterns, a module for the construction of a general definition catalog, and a module for the extraction of the candidate answer. The following sections describe in detail these modules.

This method was specially suited for answering definition questions as specified in the CLEF. That is, questions asking for the position of a person, e.g., *Who is Vicente Fox?*, and for the description of an acronym, e.g., *What is the CERN?*. Nevertheless, it is important to mention that this method was general enough to be applied to other kind of definition questions such as "*What is quinoa?*".

It is also important to point out that the processes of pattern discovery and catalog construction were done offline, whereas the answer extraction was done online. The proposed method did not consider any module for document or passage retrieval.

Pattern Discovery

The module for pattern discovery used a small set of concept-description pairs to collect from the Web an extended set of definition instances. Then, it applied a text mining method on the collected instances in order to discover a set of definition surface patterns.

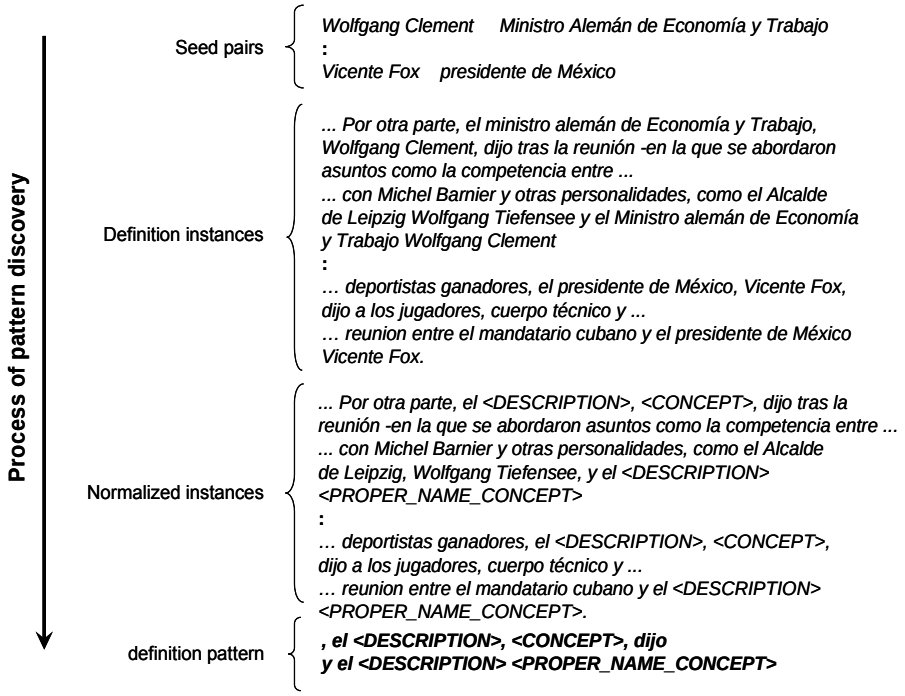

Fig. 2. Data flow in the pattern discovery process

Figure 2 illustrates the information processing through the pattern discovery process. The idea was to obtain several surface definition patterns starting up with a small set of concept-description example pairs. First, using a small set of concept description seeds, for instance, "*Wolfgang Clement – German Federal Minister of Economics and Labor*" and "*Vicente Fox – President of Mexico*", we obtained a set of definition instances. One example of these instances was "*...meeting between the Cuban leader and the president of Mexico, Vicente Fox.*". Then, the instances were normalized, and finally a sequence-mining algorithm was used to obtain some lexical patterns highly

[8] This method was an adaptation of that previously proposed in [11].

related to concept definitions. Figure 2 shows two example patterns: "*, the <DESCRIPTION>, <CONCEPT>, says*" and "*the <DESCRIPTION> <CONCEPT>*". Notice that the discovered patterns included words, punctuation marks as well as proper name tags as delimiting elements.

Catalog Construction

In this module, the definition patterns discovered in the previous step (i.e., in the pattern discovery module) were applied on the target document collection. The result was a set of matched text segments that presumably contain a concept and its description. The definition catalog was created gathering all matched segments.

Answer Extraction

Figure 3 shows the process of answer extraction for the question "*Who is Diego Armando Maradona?*". First, we obtained all descriptions associated to the requested concept. It is clear that there were erroneous or incomplete descriptions (e.g. "*Argentina soccer team*"). However, most of them contained a partially satisfactory explanation of the concept. Actually, we detected correct descriptions such as "*captain of the Argentine soccer team*" and "*Argentine star*". Then, a mining process allowed detecting a set of maximal frequent sequences. Each sequence was considered a candidate answer. In this case, we detected three sequences: "*argentine*", "*captain of the Argentine soccer team*" and "*overuse of Ephedrine by the star of the Argentine team*". Finally, the candidate answers were ranked based on the frequency of occurrence of its subsequences in the whole description set. In this way, we took advantage of the incomplete descriptions of the concept. The selected answer was "*captain of the Argentine national football soccer team*", since it was conformed from frequent subsequences such as "*captain of the*", "*soccer team*" and "*Argentine*".

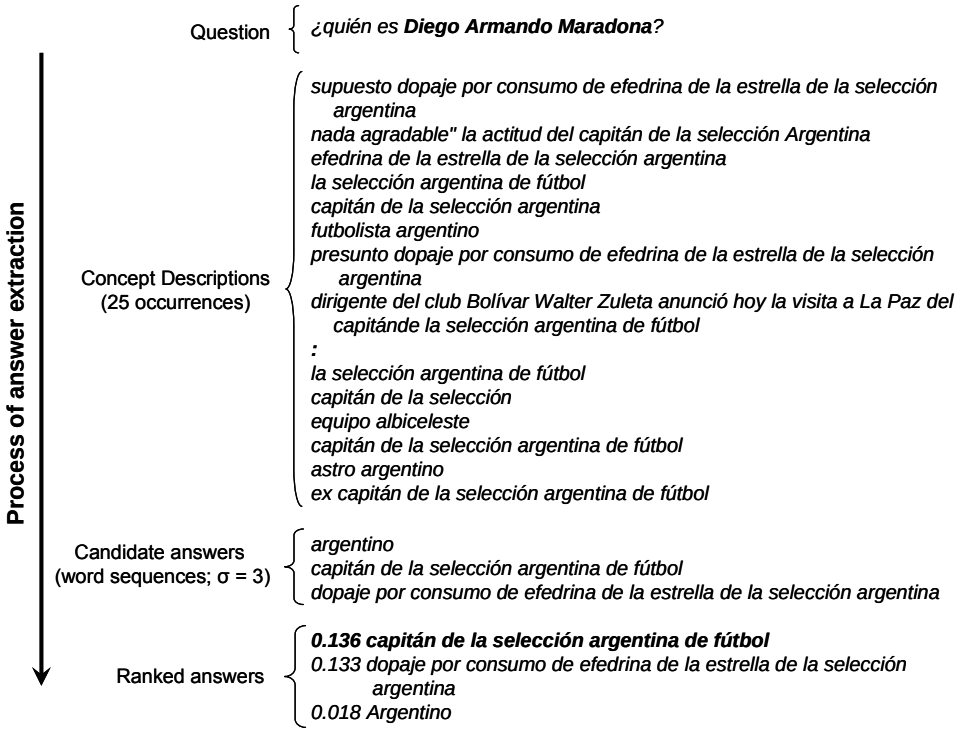

Fig. 3. Data flow in the answer extraction process

4.3.1.2 Evaluation results

Table 8 shows the evaluation results of our system at QA@CLEF 2006. Despite the fact that our results (for factual questions) were only 2% better than previous year, we believe that the proposed approach could be a good starting point for the introduction of syntactic information into the answer selection process.

Table 8. Results of system 1 at QA@CLEF 2006

	No. questions	Right	Accuracy
Definition	42	35	83.3%
Factoid	158	45	28.5%
TOTAL	200	80	40.0%

In relation to the processing of definition questions, these results indicate that our new method was very pertinent (obtaining the best accuracy rate from all participants). Different from our previous year approach, that only used 2 general definition patterns, this year we could use a much bigger set of –particular– patterns, and consequently we could find descriptions for many more persons and organizations.

4.3.2 System 2: Supervised data-driven QA

This second prototype also continued our previous year work (refer to section 4.2.2) by implementing a lexical full data-driven approach. However, it presented an important modification: it applied a supervised machine learning approach instead of a statistical method for answering factoid questions [22]. In particular, the use of machine learning allowed us considering a greater number of characteristics for building the answer extraction model.

Following, we describe the new answer extraction process; the module for answering definition questions is the same that the one described in section 4.3.1.1.

4.3.2.1 System components

The answer extraction process was based on a supervised machine learning approach. It consisted of two main modules, one for attribute extraction and another for answer selection.

Attribute extraction

In a first step, the set of retrieved passages were processed with the aim of identifying all text fragments related to the expected answer type. This process was done using a set of regular expressions that allows identifying proper names, dates, and quantities. Each identified text fragment was considered a "candidate answer".

In a second step, the lexical context of each candidate answer was analyzed with the aim of constructing its formal representation. In particular, each candidate answer was represented by a set of 17 attributes, clustered in the following groups:

1. Attributes that describe the complexity of the question. For instance, the length of the question, the number of non-stop words, etc.
2. Attributes that measure the similarity between the context of the candidate answer and the given question. Basically, these attributes consider the number of common words, word lemmas, and named entities (proper names) between the context of the candidate answer and the question. They also take into consideration the density of the question words in the answer context.
3. Attributes that indicate the relevance of the candidate answer according to the set of retrieved passages. For instance, the relative position of passage that contains the candidate answer as well as the redundancy of the answer in the whole set of passages.

Answer Selection

This module calculated for each candidate answer its probability for being the correct answer. In order to do that, it first constructed a probabilistic answer extraction model from a given training corpus. It used the Naïve Bayes classifier [44] and the sets of questions and documents from previous CLEF campaigns for building this model.

Then, in a second step, it applied the resulting model to all candidate answers of a given question, and determined their probability of being a correct answer. Finally, it selected the answer with the greatest probability as the final response.

4.3.2.2 Evaluation results

Table 9 shows the overall results corresponding to the evaluation of our full data-driven system. The system obtained excellent results. It went beyond the barrier of the 50% of accuracy and was 10-points over our previous year results (refer to section 4.2.2.2).

The method for answer extraction of factoid questions achieved good results; however, it had two significant disadvantages: (i) it required a lot of training data, and (ii) the detection of the candidate answers was not always (neither for all cases nor for all languages) an easy –high precision– task.

Table 9. Results of system 2 at QA@CLEF 2006

	No. questions	Right	Accuracy
Definition	42	35	83.3%
Factoid	158	67	42.4%
TOTAL	200	102	51.0%

On the other hand, besides the outstanding results of our method for answering definition questions, it was surprising that it replied very exact answers most of the times. Nevertheless, it had the inconvenience of constructing an enormous definition catalog (1,772,918 for acronyms and 3,525,632 for persons descriptions) containing a huge quantity of incorrect/incomplete registers.

4.4 CLEF 2007

For this year we decided to exclusively develop a full data-driven QA approach [37]. This decision was mainly supported on the fact that our knowledge-based approach was very sensible to the deficiencies of the syntactic parser, and therefore, it must be completely redesign. In addition, our full data-driven approach could be easily adapted to the new characteristics of the Spanish QA@CLEF 2007 evaluation data: the search collection was enlarged by the inclusion of Wikipedia pages, and there were also included some contextually-related questions, i.e., groups of related questions, where the first one indicates the focus of the group and the rest of them are somehow dependent from it.

4.4.1 System overview

In 2007 we presented a straightforward QA approach that avoided using any kind of linguistic resource, and therefore, that could be –in theory– applied to answer questions in several languages. This system continued our work of 2006 (refer to section 4.3.2), but incorporated some new elements. The answer extraction module was the same from our previous year prototype; however, the passage retrieval module changed to deal with the new conditions of the evaluation. On the other hand, the definition extraction component took advantage of the structure of the document collection (Wikipedia, in this case) to easily locate definition phrases. The following subsections describe these modifications.

4.4.1.1 Passage retrieval

This module aimed, as we previously mentioned, to obtain a set of relevant passages from all target document collections, in this particular case, the EFE news collection and the Spanish Wikipedia. It was primary based on a traditional Vector Space Model, but also incorporated a novel query expansion technique. This technique was mainly used to enhance the retrieval of information from Wikipedia: it allowed to deal with the low redundancy of the target collection by searching relevant information through related concepts.

The passage retrieval module considered four main processes: association rule mining, query generation, passage retrieval and passage integration. Following we describe each one of these processes.

Association rule mining

This process was done offline. Its purpose was to obtain all pairs of highly related concepts (i.e., named entities) from a given document collection. It considered that a concept A was related or associated to some other concept B (i.e., A → B), if B occurred in σ% of the documents that contains A.

In order to discover all association rules satisfying a specified σ-threshold, this process applied the well-known Apriori algorithm [2]. Using this algorithm it was possible to discover association rules such as "*Churchill → Second World War*" and "*Ernesto Zedillo → Mexico*".

Query generation

This process used the discovered association rules to automatically expand the input question. Basically, it constructed four different queries from the original question. The first query was the set of keywords (for instance, the set of named entities) from the original question, whereas the other three queries expanded the first one by including some concepts associated with all keywords of the given question. For instance, given a question such as "*Who was the president of Mexico during the Second World War?*", this process generated the following four queries: (1) "*Mexico Second World War*", (2) "*Mexico Second World War 1945*", (3) "*Mexico Second World War United States*", and (4) "*Mexico Second World War Pearl Harbor*".

Passage integration

The first step of this process consisted of recovering the largest number of relevant passages from the target document collections (EFE and Wikipedia). In order to do that, it retrieved passages using all generated queries[9]. Later, it combined the retrieved passages into one single set. Its objective was to sort all passages in accordance with a homogeneous weighting scheme. The new weight of passages was calculated as follows:

$$w_p = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{|G_i|}\sum_{y\in G_i} C(p,y)\right) \tag{9}$$

Where w_p is the new weight of passage p, n indicates the number of words of the reference question, G_i is the set of all n-grams of size i from the question, and $C(p,y)$ is 1 if the question n-gram y occurs in the passage p, otherwise it is 0. This new weighting scheme favored those passages sharing the highest number of n-grams with the question.

4.4.1.2 Answering contextually-related questions

As we previously mentioned, the 2007 evaluation included a new challenge: groups of related questions, where the first one indicates the focus of the group and the rest of them are somehow dependent on it. For instance, the pair of questions "*When was Amintore Fanfani born?*", and "*where was he born?*".

Our approach for answering this kind of questions was quite simple. It basically considered the enrichment of dependent questions by adding some keywords as well as the answer from the first (head) question. Our first idea was to resolve the anaphora and add the referent to all dependent questions. Nevertheless, given that it is impossible to assure a correct answer for the first question, we decided to add all possible elements to dependent questions, and in this way, increment the probabilities for their adequate treatment.

The process for answering a list of related questions was as follows:

1. Handle head questions as usual (refer to previous sections).
2. Extract the set of keywords (in our case, the set of named entities) from the head question. This process is done using a set of regular expressions.
3. Add to all dependent questions the set of keywords and the extracted answer from the head question.
4. Handle enriched dependent questions as usual (refer to previous sections).

[9] This process was carried out by the Lucene information retrieval system.

For instance, after this process the example question "*where was he born?*" was transformed to the enriched question "*where he was born + Amintore Fanfani + 6 February 1908*".

4.4.1.3 Answering definition questions

Our method for answering definition questions used Wikipedia as target document collection. It took advantage of two known facts: (i) Wikipedia organizes the information by topics, that is, each document concerns one single subject and, (ii) the first paragraph of each document tends to contain a short description of the topic at hand. This way, it simply retrieved the document(s) describing the target term of the question and then returned some part of its initial paragraph as answer. The general process for answering definition questions consisted of two main modules: document retrieval and answer extraction. The following sections briefly describe these modules.

Finding Relevant Documents

In order to search in Wikipedia for the most relevant document to the given question, it was necessary to firstly recognize the target term. For this purpose our method used a set of manually constructed regular expressions such as: *"What|Which|Who|How"+"any form of verb to be"+<TARGET>+"*?", *"What is a <TARGET> used for?"*, *"What is the purpose of <TARGET>?"*, *"What does <TARGET> do?"*, etc.

Then, the extracted target term was compared against all document names and the document having the highest similarity was obtained and delivered to the answer extraction module. In particular, our system used the Lucene[10] information retrieval system for both indexing and searching.

Extracting the Target Definition

As we previously mentioned, most Wikipedia documents tend to contain a brief description of its topic in the first paragraph. Based on this fact, our method for answer extraction was defined as follows:

1. Consider the first sentence of the retrieved document as the target definition (the answer).
2. Eliminate all text between parenthesis (the goal is to eliminate comments and less important information).
3. If the constructed answer is shorter than a given specified threshold[11], then aggregate as many sentences of the first paragraph as necessary to obtain an answer of the desired size.

For instance, the answer for the question "*Who was Hermann Emil Fischer?*" was extracted from the first paragraph of the document "Hermann Emil Fischer". Below we illustrate this paragraph, where underline words indicate the selected definition.

> "Hermann Emil Fischer (October 9, 1852 – July 15, 1919) was a German chemist and recipient of the Nobel Prize for Chemistry in 1902. Emil Fischer was born in Euskirchen, near Cologne, the son of a businessman. After graduating he wished to study natural sciences, but his father compelled him to work in the family business until determining that his son was unsuitable".

4.4.2 Evaluation results

Table 10 details our general accuracy results. It is interesting to notice that our method for answering definition questions was very precise. It could answer 87.5% of the questions; moreover, it never replied wrong or unsupported answers. This result evidenced that Wikipedia has an inherent structure, and that our method could effectively take advantage of it. On the other hand, Table 7 also shows that our method for answering factoid questions was not completely adequate (it only could answer 24% of this kind of questions). Taking into consideration that this method obtained 42% of accuracy on previous year exercise (4.3.2.2), we suspect that its poor performance was caused by the inclusion of Wikipedia. Two characteristics of Wikipedia damaged our system behavior. First, it is much less redundant that general news collections; and second, its style and structure make lexical contexts of candidate answers less significant that those extracted from other free-text collections.

10 lucene.apache.org

11 For the experiments reported we defined this threshold equal to 70 non-blank characters. This number was estimated after a manual analysis of several Wikipedia documents.

Table 10. Results of our system at QA@CLEF 2007

	No. questions	Right	Accuracy
Definition	32	28	87.5%
Factoid	168	41	24.4%
TOTAL	200	69	34.5%

Finally, Table 11 shows the results about the treatment of contextually-related questions. It is clear that the proposed approach (refer to section 4.4.1.2) was not useful for dealing with contextually-related questions. The reason of this poor performance is that only 37% of head questions were correctly answered, and therefore, in the majority of the cases dependent questions were enriched with erroneous information.

Table 11. Evaluation details about answering dependent questions

	No. questions	Right	Accuracy
Head questions	170	64	37.65%
Dependent questions	30	5	16.67%

5 Conclusions and Perspectives

This work had as objective to show the Mexican contributions to the solution of the problems underlying QA in Spanish. In particular, we focused on the works evaluated in the forum QA@CLEF carried out in the last years by the Laboratory of Language Technologies (LabTL) of INAOE. In this forum, several research groups propose and evaluate their ideas for the search of answers. The forum formulates an objective evaluation utilizing the same conditions and criteria. Along these years, the LabTL developed several systems that demonstrated their effectiveness when compared against those proposed by other research groups, mainly from Europe - in fact, the LabTL has been the only Latin American team participating in CLEF, in these last 4 years. Thanks to this active participation in CLEF, the LabTL has consolidated this line of research, and its works have received wide recognition from the rest of participants in CLEF. It is worth to mention that the proposed methods are among the best ranked since 2005, getting ahead of most of the participating research groups, except of a private company.

The systems proposed by LabTL can be grouped in two big approaches: the first, using a purely statistical conceptualization, language independent, and not requiring linguistic tools; and the second, a linguistic approach that attempts to solve the problem of QA by recurring to the underlying structures of Spanish. Despite of the results obtained so far, none of the two approaches has reached adequate levels for their practical use. Among the main disadvantages we can mention the following. Regarding the first approach, it is necessary to apply on huge collections, where the probabilities of finding the answer expressed in the same terms as the question are increased, and additionally the answer is mentioned repeatedly. This restriction aggravates when the source of information is semi-structured, as is the case of Wikipedia. With respect to the second approach, the problem is centered on the availability of the linguistic tools used. Moreover, the demand required from existing tools is very high, pretending to work on open domains. In this way, the low performance of linguistic tools causes an effect of cascade errors, diminishing strongly the final precision.

An alternative formulation, also experienced by LabTL, is the search of a scheme that combines these two approaches. This consists of including a module to recognize the textual implication (TIR). In this case, the idea is using the language independent approach in order to select a small set of candidate answers, which are further examined by the TIR module to pick the correct answer, if this exists in the given set. The module of TIR [38] makes its selection after carrying out a linguistic analysis of the question and of the passage where the candidate answer appears. In this way, it is possible to combine the successes of both approaches.

Finally, the works carried out by the LabTL have opened the doors to multilingual information processing. With our language independent approach, evaluated even in French and Italian at CLEF 2005, it is possible to deepen in

the search of answers in multilingual collections. This is highly relevant when the Web is seen as a source of information, where there exist enormous quantities of information in multiple languages. So far, it is unquestionable the predominance of English in the Web, but it is also clear the increment of other languages, Spanish among them. This motivates the growing need of automatic tools that allow the access, in simple and practical way, to this huge repository of multilingual information. We have made in Mexico a noteworthy contribution to the state of the art in QA in this direction.

Acknowledgements

The authors want to recognize the direct contribution of doctoral students Alberto Téllez-Valero, Manuel Pérez-Coutiño, Rita M. Aceves-Pérez, and master students Antonio Juárez-González, Claudia Denicia-Carral, as well as the indirect efforts of doctoral students Thamar Solorio, René García-Hernández, and master students Gustavo Hernández-Rubio, Esaú Villatoro-Tello, Aarón Pancardo-Rodríguez, and Ma. del Rosario Peralta-Calvo. This research was partially supported by Conacyt through research grants C01-39957, 43990A-1, and SNI.

References

1. **Ageno, A., Ferrés, D., González, E., Kanaan, S., Rodríguez H., Surdeanu, M., and Turmo J.** *TALP-QA System for Spanish at CLEF-2004*. In CLEF 2004 Working Notes. Bath, UK. September 2004.
2. **Agrawal, R., and Srikant, R.** *Fast Algorithms for Mining Association Rules*. Proceedings of the 20th. VLDB Conference. Santiago de Chile, Chile. 1994.
3. **Amaral, C., Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., and Vidal, D.** *Priberam's Question Answering System in QA@CLEF 2007*. In CLEF 2007 Working Notes. Budapest, Hungary. September 2007.
4. **Aunimo, L., and Kuuskoski, R.** *Question Answering using Semantic Annotation*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.
5. **Bertagna, F., Chiran, L., and Simi, M.** *QA at ILC-UniPi: Description of the Prototype*. In CLEF 2004 Working Notes. Bath, UK. September 2004.
6. **Brill, E., Lin, J., Banko, M., Dumais, S., and Ng, A.,** *Data-intensive Question Answering*. In TREC 2001 Proceedings. Maryland, USA. November 2001.
7. **Buscaldi, D., Gómez, J. M., Rosso, P., and Sanchis, E.** *The UPV at QA@CLEF 2006*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
8. **Buscaldi, D., Benajiba, Y., Rosso P. and Sanchis. E.** *The UPV at QA@CLEF 2007*. In CLEF 2007 Working Notes. Budapest, Hungary. September, 2007.
9. **Cassan, A., Figueira, H., Martins, A., Mendes, A., Mendes, P., Pinto, C., and Vidal, D.** *Priberam's Question Answering System in a Cross-Language Environment*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
10. **Del-Castillo, A., Montes-y-Gómez, M., and Villaseñor-Pineda, L.** *QA on the web: A preliminary study for Spanish language*. In Proceedings of the 5th Mexican International Conference on Computer Science (ENC04), Colima, Mexico. September 2004.
11. **Denicia-Carral, C., Montes-y-Gómez, M., Villaseñor-Pineda, L., and García-Hernández, R.** *A Text Mining Approach for Definition Question Answering*. In Proceedings for the Fifth International Conference on Natural Language Processing, FinTAL 2006. Turku, Finland. August 2006.
12. **de-Pablo-Sánchez, C., Martínez-Fernández, J.L. , Martínez, P., Villena, J., García-Serrano, A.M., Goñi, J.M. and González, J.C.** *miraQA: Initial Experiments in Question Answering*. In CLEF 2004 Working Notes. Bath, UK. September 2004.
13. **de-Pablo-Sánchez, C., González-Ledesma, A., Martinez-Fernández, J.L., Guirao, J.M., Martínez, P., and Moreno, A.,** *MIRACLE's 2005 Approach to Cross-Lingual Question Answering*, In CLEF 2005 Working Notes. Vienna, Austria. September 2005.

14. **de-Pablo-Sánchez, C., González-Ledesma, A., Moreno, A., Martínez-Fernández, J. L., and Martínez, P.,** *MIRACLE at the Spanish CLEF@QA 2006 Track*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
15. **de-Pablo-Sánchez, C. Martínez, J.L., García-Ledesma, A., Samy, D., Martínez, P., Moreno-Sandoval A., and Al-Jumaily, H.** *MIRACLE Question Answering System for Spanish at CLEF 2007*. In CLEF 2007 Working Notes. Budapest Hungary. September, 2007.
16. **Ferrández, S., López-Moreno, P., Roger, S.,. Ferrández, Peral, J., Alvarado, X., Noguera, E., and Llopis, F.** *AliQAn and BRILI QA Systems at CLEF 2006*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
17. **Ferrés, D. Kanaan, S., González, E., Ageno, A., Rodríguez, H., and Turmo, J.** *The TALP-QA System for Spanish at CLEF-2005*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.
18. **García-Cumbreras, M. A., Ureña-López, L. A., Martínez-Santiago, F., and Perea-Ortega, J. M.** *BRUJA System. The University of Jaén at the Spanish Task of CLEFQA 2006*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
19. **Giampiccolo, D., Forner, P., Peñas, A., Ayache, C., Cristea, D., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., and Sutcliffe, R.** *Overview of the CLEF 2007 Multilingual Question Answering Track*. In CLEF 2007 Working Notes. Budapest, Hungary. September 2007.
20. **Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., and Rosso, P.** *A Passage Retrieval System for Multilingual Question Answering*. In Proceedings of the 8th International Conference on Text, Speech and Dialog, TSD 2005. Karlovy Vary, Czech Republic, September 2005.
21. **Gómez-Soriano, J.M., Bisbal-Asensi, E., Buscaldi, D., Rosso, P., and Sanchís, E.** *Monolingual and Cross-language QA using a QA-oriented Passage Retrieval System*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.
22. **Juárez-González, A., Téllez-Valero, A., Denicia-Carral, C., Montes-y-Gómez, M., and Villaseñor-Pineda, L.** *INAOE at CLEF 2006: Experiments in Spanish Question Answering*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
23. **Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo F., and de Rijke M.** *The Multiple Language Question Answering Track at CLEF 2003*. In CLEF 2003 Workshop Notes. Trondheim, Norway. August 2003.
24. **Magnini, B., Vallin A., Ayache C., Erbach G., Peñas A., de Rijke M., Rocha P., Simov K. and Sutcliffe R.** *Overview of the CLEF 2004 Multilingual Question Answering Track*. In CLEF 2004 Working Notes. Bath, UK. September 2004.
25. **Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Osenova, P., Peñas, A., Jijkoun, V., Sacaleanu, B., Rocha, P., and Sutcliffe, R.** *Overview of the CLEF 2006 Multilingual Question Answering Track*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
26. **Méndez-Díaz, E., Vilares-Ferro, J. and Cabrero-Souto, D.** *COLE at CLEF 2004: Rapid Prototyping of a QA system for Spanish*. In CLEF 2004 Working Notes, Bath, UK, September 2004.
27. **Montes-y-Gómez, M., Villaseñor-Pineda, L., Pérez-Coutiño, M., Gómez-Soriano, J.M., Sanchis-Arnal, E. and Rosso, P.** *INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.
28. **Pérez-Coutiño, M., Solorio, T., Montes-y-Gómez, M., López-López, A., and Villaseñor-Pineda, L.** *The Use of Lexical Context in Question Answering for Spanish*. In CLEF 2004 Working Notes, Bath, UK. September 2004.
29. **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 Proceedings of the Second International Atlantic Web Intelligence Conference. AWIC 2004. Cancun, Mexico. May 2004.
30. **Pérez-Coutiño, M., Montes-y-Gómez, M., López-López, A., and Villaseñor-Pineda L.** *Experiments for Tuning the Values of Lexical Features in Question Answering for Spanish*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.

31. **Pérez-Coutiño, M., Montes-y-Gómez, M. López-López, A., Villaseñor-Pineda, L., and Pancardo-Rodríguez, A.** *A Shallow Approach for Answer Selection based on Dependency Trees and Term Density*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
32. **Prager J., Radev D., Brown E., Coden A. and Samn V.** *The Use of Predictive Annotation for Question Answering in TREC8*. In TREC 8 Proceedings. Maryland, USA. November 1999.
33. **Ravichandran, D., and Hovy, E.** *Learning Surface Text Patterns for a Question Answering System*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, USA. July 2002.
34. **Roger S., Ferrández S., Ferrández A., Peral J., Llopis F., Aguilar, A. and Tomás D.** *AliQAn, Spanish QA System at CLEF-2005*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.
35. **Saggion, H.** *Identifying Definitions in Text Collections for Question Answering*. The fourth international conference on Language Resources and Evaluation, LREC 2004. Lisbon, Portugal. May 2004.
36. **Tanev, H., Kouylekov, M., Magnini B., Negri, M., and Simov, K.,** *Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-irst at CLEF 2005*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.
37. **Téllez-Valero, A., Juárez-González, A., Hernández-Rubio, G., Delicia-Carral, C., Villatoro-Tello, E., Montes-y-Gómez, M., and Villaseñor-Pineda, L.** *INAOE's Participation at QA@CLEF 2007*. In CLEF 2007 Working Notes. Budapest Hungary, September, 2007.
38. **Téllez-Valero, A., Montes-y-Gómez, M., and Villaseñor-Pineda, L.** *INAOE at AVE 2007: Experiments in Spanish Answer Validation*. In CLEF 2007 Working Notes. Budapest, Hungary. September 2007.
39. **Tomás, D., Vicedo, J. L., Saiz, M., and Izquierdo, R.** *Building an XML Framework for Question Answering*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.
40. **Tomás, D., Vicedo, J. L., Bisbal, E., and Moreno, L.** *Experiments with LSA for Passage Re-Ranking in Question Answering*. In CLEF 2006 Working Notes. Alicante, Spain. September 2006.
41. **Vallin A., Giampiccolo D., Aunimo L., Ayache C., Osenova P., Peñas A., de Rijke M., Sacaleanu B., Santos D., and Sutcliffe R.** *Overview of the CLEF 2005 Multilingual Question Answering Track*. In CLEF 2005 Working Notes. Vienna, Austria. September 2005.
42. **Vicedo, J.L., Izquierdo, R., Llopis, F., and Muñoz, R.,** *Question Answering in Spanish*., In CLEF 2003 Workshop Notes. Trondheim, Norway. August 2003.
43. **Vicedo, J. L., Saiz, M., and Izquierdo, R.** *Does English help Question Answering in Spanish?* In CLEF 2004 Working Notes, Bath, UK. September 2004.
44. **Witten H. and Frank E.** *Data Mining: Practical Machine Learning Tools and Techniques*. Second edition. Morgan Kaufmann. 2005.

Manuel Montes y Gómez. *He is a full-time lecturer at the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. He obtained his PhD in Computer Science from the Computing Research Center of the National Polytechnic Institute, Mexico. His research interests include question answering, information extraction, information retrieval, text classification, and text mining.*

Luis Villaseñor Pineda. *He obtained his PhD in Computer Science from the Université Joseph Fourier, Grenoble, France. Since 2001, he is professor of the Department of Computational Sciences of the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. His areas of interest include questions answering, thematic and non-thematic text classification, and speech recognition and dialogue systems.*

Aurelio López López *is a professor at the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. He got his PhD in Computer and Information Science from Syracuse University, Syracuse NY, U.S.A. His areas of interest include knowledge representation, information extraction and retrieval, information processing, natural language processing, and text mining.*