A Text Mining Approach for Definition Question Answering

Claudia Denicia-Carral, Manuel Montes-y-Gómez,
Luis Villaseñor-Pineda, René García Hernández

Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{cdenicia, mmontesg, villasen, renearnulfo}@inaoep.mx

Abstract. This paper describes a method for definition question answering based on the use of surface text patterns. The method is specially suited to answer questions about person's positions and acronym's descriptions. It considers two main steps. First, it applies a sequence-mining algorithm to discover a set of definition-related text patterns from the Web. Then, using these patterns, it extracts a collection of concept-description pairs from a target document database, and applies the sequence-mining algorithm to determine the most adequate answer to a given question. Experimental results on the Spanish CLEF 2005 data set indicate that this method can be a practical solution for answering this kind of definition questions, reaching a precision as high as 84%.

1 Introduction

Nowadays, thanks to the Internet explosion, there is an enormous volume of available data. This data may satisfy almost every information need, but without the appropriate search facilities, it is practically useless. This situation motivated the emergence of new approaches for information retrieval such as question answering.

A question answering (QA) system is an information retrieval application whose aim is to provide inexperienced users with a flexible access to information, allowing them writing a query in natural language and obtaining not a set of documents that contain the answer, but the concise answer itself [11]. At present, most QA systems focus on treating short-answer questions such as factoid, definition and temporal. This paper focuses on answering definition questions as delimited in the CLEF[1]. These questions, in contrast to those of TREC[2], exclusively ask for the position of a person, e.g., Who is George Bush?, and for the description of an acronym, e.g., What is UNICEF?.

There are several approaches to extract answers from free text for this kind of questions. Most of them take advantage of some stylistic conventions frequently used by writers to introduce new concepts. These conventions include some typographic elements that can be expressed by a set of lexical patterns. In the initial attempts, these patterns were manually created [5, 9]. However, because they are difficult to

[1] Cross-Language Evaluation Forum (www.clef-campaign.org).
[2] Text REtrieval Conference (trec.nist.gov/)

extract and domain dependent, current approaches tend to construct them automatically [2, 8].

In this paper, we explore a text mining approach for answering this kind of definition questions. In particular, we use a sequence-mining algorithm [1] to discover definition patterns from the Web as well as to identify the best candidate answer to a given question from a set of matched concept-description pairs. The double use of the sequence-mining algorithm gives our method its power. It allows the discovery of surface definition patterns for any kind of text or domain, and enables taking advantage on the redundancy of the target document collection to determine with finer precision the answers to the questions.

In order to evaluate this method, we consider the definition questions from the Spanish CLEF 2005 evaluation exercise. Our results demonstrate that our approximation can be effectively used to answer definition questions from free-text documents.

The rest of the paper is organized as follows. Section 2 discuses some related work. Section 3 presents the general scheme of the method. Section 4 describes their main components. Section 5 introduces the task of sequence mining and explains our approach to answer ranking. Section 6 shows the experimental results, and finally, section 7 presents our conclusions and future work.

2 Related Work

There are several approaches for answering definition questions. Most of them use lexical patterns to extract the answer to a given question from a target document collection. Depending on the complexity of the requested definition, it is the complexity of the useful patterns. For the simplest case, i.e., the introduction of a new referent in the discourse, the stylistic conventions used by authors are clear and stable. In consequence, the practical lexical patterns are simple and precise. Under this assumption, the questions like “What is X?” and “Who is Y?” are resolved.

The existing approaches for answering definition questions diverge in the way they determine the definition patterns and in the way they use them. There are some works that applies patterns that were manually constructed [5, 9, 3], and other works that automatically construct the patterns from a set of usage examples [2, 8]. Our method considers the automatic construction of the patterns. It consists of two main steps:

In the first step, the method applies a mining algorithm in order to discover a set of definition-related text patterns from the Web. These lexical patterns allow associating persons with their positions, and acronyms with their descriptions. This step is similar to other previous approaches (especially to [8]). Nevertheless, our method differs from them in that it considers all discovered patterns, i.e., it does not evaluate and select the mined patterns. Therefore, the main difference in this first step is that while others focus on selecting a small number of very precise patterns, we concentrate on discovering the major number of mutually exclusive patterns.

In the second step, the method applies the patterns over a target document collection in order to answer the specified questions. The way we use the patterns to answer definition questions is quite novel. Previous works [8, 5, 9] apply the patterns over a set of “relevant” passages, and trust that the best (high-precision) patterns will

allow identifying the answer. In contrast, our method applies all discovered patterns to the entire target document collection and constructs a "general catalog". Then, when a question arrives, it mines the definition catalog in order to determine the best answer for the given question. In this way, the answer extraction does not depend on a passage retrieval system and takes advantage on the redundancy of the entire collection.

3 Method at a Glance

Figure 1 shows the general scheme of our method. It consists of two main modules; one focuses on the discovery of definition patterns and the other one on the answer extraction.

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

The module for answer extraction applies the discovered patterns over a target document collection in order to create a definition catalog consisting of a set of potential concept-description pairs. Later, given a question, it extracts from the catalog the set of associated descriptions to the requested concept. Finally, it mines the selected descriptions to find the more adequate answer to the given question.

It is important to notice that the process of pattern discovery is done offline, while the answer extraction, except for the construction of the definition catalog, is done

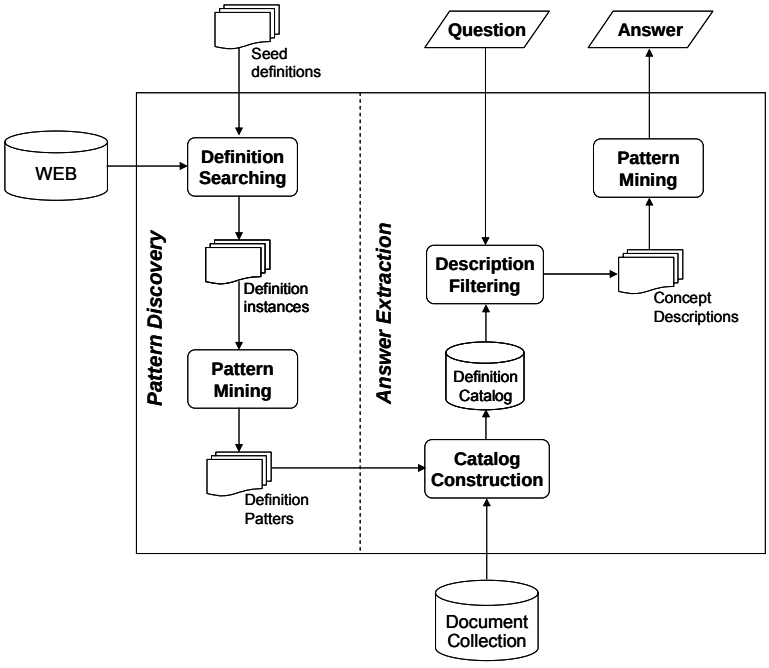

Figure 1. General diagram of the method

online. It is also important to mention that different to traditional QA approaches, the proposed method does not consider any module for document or passage retrieval. The following section describes in detail these two modules.

4 Answering Definition Questions

4.1 Pattern Discovery

As we mentioned, there are certain stylistic conventions frequently used by authors to introduce new concepts in a text. Several QA approaches exploit these conventions by means of a set of lexical patterns. Unfortunately, there are so many ways in which concepts are described in natural language that it is difficult to come up with a complete set of linguistics patterns to solve the problem. In addition, these patterns depend on the text domain, writing style and language.

In order to solve these difficulties we use a very general method for pattern discovery [8]. The method captures the definition conventions through their repetition. It considers two main subtasks:

Definition searching. This task is triggered by a small set of empirically defined concept-description pairs. The pairs are used to retrieve a number of usage examples from the Web[3]. Each usage example represents a definition instance. To be relevant, a definition instance must contain the concept and its description in one single phrase.

Pattern mining. It is divided in three main steps: data preparation, data mining and pattern filtering.

The purpose of the data preparation phase is to normalize the input data. In this case, it transforms all definition instances into the same format, using special tags for the concepts and their descriptions.

In the data mining phase, a sequence mining algorithm (refer to section 5.1) is used to obtain all maximal frequent sequences –of words and punctuation marks– from the set of definition instances. The sequences express lexicographic patterns highly related to concept definitions.

Finally, the pattern-filtering phase allows choosing the more discriminative patterns. It selects the patterns satisfying the following general regular expressions:

<left-frontier-string> DESCRIPTION <center-string> CONCEPT < right-frontier-string>

<left-frontier-string> CONCEPT <center-string>DESCRIPTION < right-frontier-string>

Figure 2 illustrates the information treatment through the pattern discovery process. The idea is 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 is "*...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 lexico-

[3] At present, we are using Google for searching the Web.

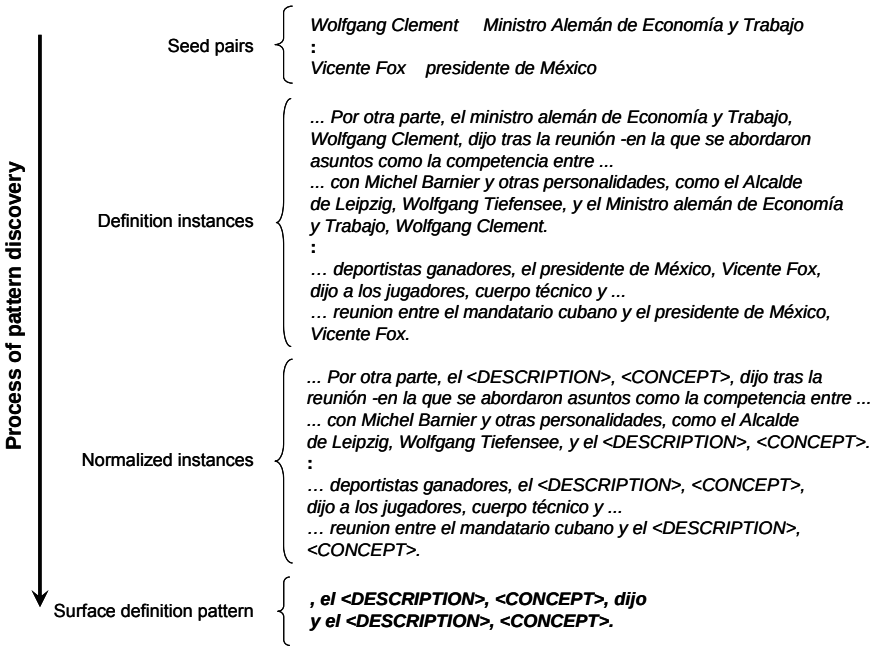

Figure 2. Data flow in the pattern discovery process

graphic patterns highly related to concept definitions. The figure shows two obtained patterns: "*, the <DESCRIPTION>, <CONCEPT>, says*" and "*and the <DESCRIPTION>, <CONCEPT>*.". It is important to notice that the patterns not only include words as frontier elements but also punctuation marks.

4.2 Answer Extraction

This second module handles the extraction of the answer for a given definition question. It is also based on a text mining approach. Its purpose is to find the more adequate description for a requested concept from an automatically constructed definition catalog.

Because the definition patterns guide the construction of the definition catalog, it contains a huge diversity of information, including incomplete and incorrect descriptions for many concepts. However, it is expected that the correct information will be more abundant than the incorrect one. This expectation supports the idea of using a text mining technique to distinguish between the adequate and the improbable answers to a given question.

This module considers the following steps:

Catalog construction. In this phase, the definition patterns discovered in the previous stage (i.e., in the pattern discovery module) are applied over the target document collection. The result is a set of matched segments that presumably contain a concept and its description. The definition catalog is created gathering all matched segments.

Description filtering. Given a specific question, this procedure extracts from the definition catalog all descriptions corresponding to the requested concept. As we mentioned, these "presumable" descriptions may include incomplete and incorrect information. However, it is expected that many of them will contain, maybe as a substring, the required answer.

Answer mining. This process aims to detect a single answer to the given question from the set of extracted descriptions. It is divided in three main phases: data preparation, data mining and answer ranking.

The data preparation phase focuses on homogenizing the descriptions related to the requested concept. The main action is to convert these descriptions to a lower case format.

In the data mining phase, a sequence mining algorithm (refer to section 5.1) is used to obtain all maximal frequent word sequences from the set of descriptions. Each sequence indicates a candidate answer to the given question.

Then, in the answer raking phase, each candidate answer is evaluated according to the frequency of occurrence of its subsequences. The idea is that a candidate answer assembled from frequent subsequences has more probability of being the correct answer than one formed by rare ones. Therefore, the sequence with the greatest ranking score is selected as the correct answer. The section 5.2 introduces the ranking score.

Figure 3 shows the process of answer extraction for the question "*Who is Diego Armando Maradona?*". First, we obtained all descriptions associated with the requested concept. It is clear that there are erroneous or incomplete descriptions (e.g. "*Argentina soccer team*"). However, most of them contain 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 "*supposed 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*".

It is important to clarify that a question may have several correct answers. In accordance with the CLEF, an answer is correct if there is a passage that supports it. Therefore, for the question at hand there are other correct answers such as "*ex capitán de la selección argentina de futbol*" and "*astro argentino*".

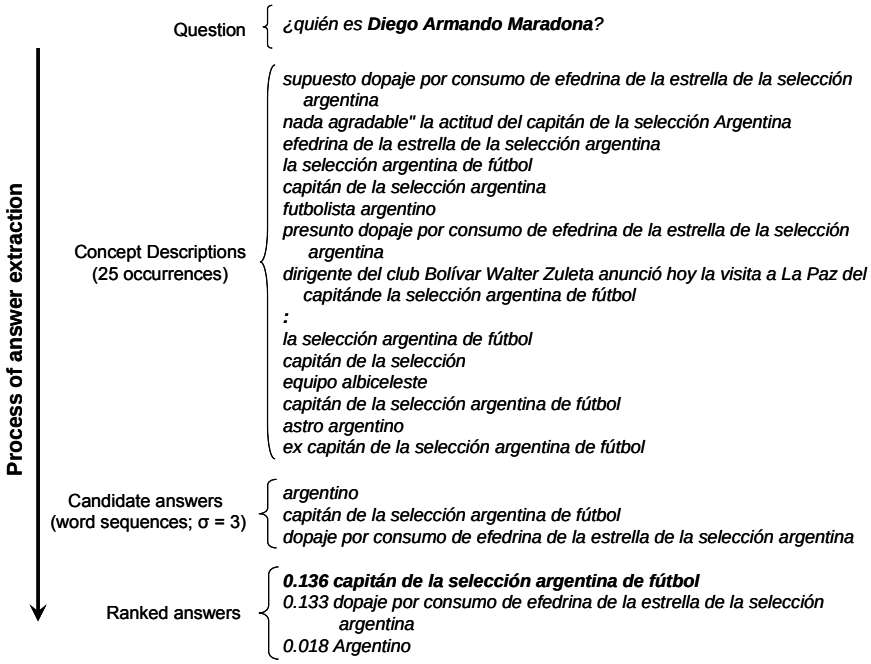

Figure 3. Data flow in the answer extraction process

5 Text mining techniques

5.1 Mining Maximal Frequent Word Sequences

Assume that D is a set of texts (a text may represent a complete document or even just a single sentence), and each text consists of a sequence of words. Then, we have the following definitions [1].

Definition 1. A sequence $p = a_1 \ldots a_k$ is a *subsequence* of a sequence q if all the items a_i, $1 \leq i \leq k$, occur in q and they occur in the same order as in p. If a sequence p is a subsequence of a sequence q, we also say that p occurs in q.

Definition 2. A sequence p is *frequent* in D if p is a subsequence of at least σ texts of D, where σ is a given frequency threshold.

Definition 3. A sequence p is a *maximal frequent sequence* in D if there does not exist any sequence p' in D such that p is a subsequence of p' and p' is frequent in D.

Once introduced the maximal frequent word sequences, the problem of mining maximal frequent word sequences can formally state as follows: Given a text collection D and an arbitrary integer value σ such that $1 \leq \sigma \leq |D|$, enumerate all maximal frequent word sequences in D.

The implementation of a method for sequence mining is not a trivial task because of its computational complexity. The algorithm used in our experiments is described in [4].

5.2 Ranking score

This measure aims to establish the better answer for a given definition question. Given a set of candidate answers (the maximal frequent sequences obtained from the set of concept descriptions), this measure selects the final unique answer taking into consideration the frequency of occurrence of its subsequences.

The ranking score R for a word sequence indicates its compensated frequency. It is calculated as follows:

$$R_{p(n)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n-i+1} \frac{f_{p_j(i)}}{\sum_{\forall q \in S_i} f_{q(i)}} \tag{1}$$

In this formula, we have introduced the following notation for the sake of simplicity. S_i indicates the set of sequences of size i, $q(i)$ represents the sequence q of size i, $p_j(i)$ is the j-th subsequence of size i included in the sequence $p(n)$, $f_{q(i)}$ specifies the frequency of occurrence of the sequence q in the set of concept descriptions, and finally $R_{p(n)}$ indicates the compensated frequency of the sequence p.

The idea behind this ranking score is that a candidate answer assembled from frequent subsequences has more probability of being the correct answer than one formed by rare substrings. Evidently, the frequency of occurrence of the stopwords is not considered into the calculation of the ranking score.

6 Experimental results

In order to evaluate the proposed method, we considered the task of definition question answering. In particular, we contemplated questions asking about the position of a person as well as questions demanding the description of an acronym.

Table 1 shows some numbers on the process of pattern discovery. It is important to notice that using only 20 definition seed pairs we discovered 78 definition patterns related to positions and 122 related to acronyms. Some of these patterns are shown in table 2.

The quality of the discovered patterns is very diverse. Some are too specific and precise but not so applicable. Some others are too general, but guarantee a high coverage. The combined application of all of them represents a good compromise between precision and coverage, and produces the data redundancy required by the process of answer extraction.

Table 1. Statistics on the process of pattern discovery

Question Type	Seed Definitions	Collected Snippets	Maximal Frequent Sequences	Surface Definition Patterns
Positions	10	6523	875	78
Acronym	10	10526	1504	122

Table 2. Examples of definition patterns

Position related patterns	Acronym related patterns
El <DESCRIPTION>, <CONCEPT>, ha	*del <DESCRIPTION> (<CONCEPT>).*
del <DESCRIPTION>, <CONCEPT>.	*que la <DESCRIPTION> (<CONCEPT>)*
El ex <DESCRIPTION>, <CONCEPT>,	*de la <DESCRIPTION> (<CONCEPT>) en*
por el <DESCRIPTION>, <CONCEPT>.	*del <DESCRIPTION> (<CONCEPT>) y*
El <DESCRIPTION>, <CONCEPT>, se	*en el <DESCRIPTION> (<CONCEPT>)*

The evaluation of the answer extraction process was based on the Spanish CLEF05 data set. This set includes a collection of 454,045 documents, and a set of 50 definition questions related to person's positions and acronym's descriptions.

Table 3 shows some data on the process of answer extraction. It shows that initially we extracted quite a lot of "presumable" related descriptions per question. The purpose is to catch an answer for all questions, and to capture most of their occurrences. Then, using a text-mining technique, we detected just a few high-precision candidate answers (sequences) per question. It is important to point out that the number of candidate answers is smaller for the questions about acronyms than for those about person's positions. We consider this situation happened because positions are regularly expressed in several ways, while acronyms tend to have only one meaning.

Table 3. Statistics on the process of answer extraction

Question Type	Average Descriptions per Question	Average Candidate Answers per Question
Positions	633	5.04
Acronym	1352.96	1.67

Table 4 presents the overall precision results for the question answering evaluation exercise. The second column indicates the precision when the answers were extracted using only the sequence mining algorithm, i.e., when answers were defined as the most frequent sequences in the set of descriptions related to the requested concept. On the other hand, the last column shows the precision rates achieved when the answers were selected using the proposed ranking score.

Table 4. Overall results on definition question answering

	Answer Selection	
Question Type	**Most Frequent Sequence**	**Highest Ranking Score**
Positions	64%	80%
Acronym	80%	88%
Total	**72%**	**84%**

The results demonstrated that our method could be a practical solution to answer this kind of definition questions, reaching a precision as high as 84%. We consider that these results are very significant, since the average precision rate for definition questions on the CLEF 2005 edition was 48%, being 80% the best result and 0% de worst [10]. Indeed, the best result at CLEF 2005 for definition questions was achieved by other method proposed by our Lab. The main difference between these methods is that while the old one uses manually constructed patterns, the new approach applies automatically discovered patterns.

It is important to mention that our method could not determine the correct answer for all questions. This situation was mainly caused by the lack of information for the requested concepts in the definition catalog. In particular, the definition catalog does not contain any information related to six questions. For instance, we could not find any description for the organization "*Medicos sin Fronteras*" ("*Doctors Without Borders"*). This was because the discovered definition patterns only allow extracting descriptions related to acronyms but not locating descriptions related to complete organization names. In order to reduce this problem it is necessary to have more definition patterns that consider several different ways of describing a concept.

Finally, it is also important to mention that a major weakness of the proposed method is that it greatly depends on the redundancy of the target collection, and especially, on the redundancy of the searched answer. Therefore, if there is just one single occurrence of the searched answer in the whole collection, then our method will not have sufficient evidence to resolve the given question.

7 Conclusions and Future Work

In this paper, we presented a method for answering definition questions. This method considers two main tasks: the discovery of definition patterns from the Web, and the extraction of the most adequate answer for a given question. The use of a text mining technique in both tasks gives our method its power. It allows the discovery of surface definition patterns for any kind of text or domain, and enables taking advantage on the redundancy of the target document collection to determine with finer precision the answer to a question.

The method was evaluated on the definition questions from the Spanish CLEF 2005 data set. These questions ask about person's positions and acronym's descriptions. The obtained results are highly significant since they are superior to those reported in the CLEF 2005 working notes [10].

In addition, the results demonstrated that it is possible to answer this kind of definition questions without using any kind of linguistic resource or knowledge. Even more, they also evidenced that a non-standard QA approach, which does not contemplate an IR phase, can be a good scheme for answering definitions questions.

As future work, we plan to:

- Consider more types of definition questions. In particular we are interested in using this approach to answer questions about general things, for instance questions like "what is an aspirin?". In order to do that it will be necessary to extend the proposed approach to consider patterns beyond the lexical level.

- Apply the method on different languages. Since our method does not use any sophisticated tool for language analysis, we believe that it could be easily adapted to other languages. This way, we plan to work with other languages considered by the CLEF, such as Italian, French and Portuguese.
- Use the method to discover different kind of patterns. For instance, patterns related to different semantic relations (e.g. synonymy, hyperonymy, etc.).

Acknowledgements

This work was done under partial support of CONACYT (Project Grants 43990 and U39957-Y) and SNI-Mexico. We also thanks to the CLEF for the resources provided.

References

1. Ahonen-Myka H. (2002). Discovery of Frequent Word Sequences in Text Source. Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery. London, UK, 2002.
2. Cui H., Kan M., and Chua T. (2004). Unsupervised Learning of Soft Patterns for Generating Definitions from Online News. Proceedings International WWW Conference. New York, USA, 2004.
3. Fleischman M., Hovy E. and Echihabi A. (2003). Offline Strategies for Online Question Answering: Answering Question Before they are Asked. Proceedings of the ACL-2003, Sapporo, Japan, 2003.
4. García-Hernández, R., Martínez-Trinidad F., and Carrasco-Ochoa A. (2006). A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. International Conference on Computational Linguistics and text Processing, CICLing-2006. Mexico City, Mexico, 2006.
5. Hildebrandt W., Katz B., and Lin J. (2004). Answering Definition Questions Using Multiple Knowledge Sources. Proceedings of Human Language Technology Conference. Boston, USA, 2004.
6. Montes-y-Gómez M., Villaseñor-Pineda L., Pérez-Coutiño M., Gómez-Soriano J. M., Sanchis-Arnal E. and Rosso, P. (2003). INAOE-UPV Joint Participation in CLEF 2005: Experiments in Monolingual Question Answering. Working Notes of CLEF 2005. Vienna, Austria, 2005.
7. Pantel P., Ravichandran D. and Hovy E. (2004). Towards Terascale Knowledge Acquisition. Proceedings of the COLING 2004 Conference. Geneva, Switzerland, 2004.
8. Ravichandran D., and Hovy E. (2002). Learning Surface Text Patterns for a Question Answering System. Proceedings of the ACL-2002 Conference. Philadelphia, USA, 2002.
9. Soubbotin M. M., and Soubbotin S. M. (2001). Patterns of Potential Answer Expressions as Clues to the Right Answer. Proceedings of the TREC-10 Conference. Gaithersburg, 2001.
10. Vallin A., Giampiccolo D., Aunimo L., Ayache C., Osenova P., Peñas A., de Rijke M., Sacaleanu B., Santos D., and Sutcliffe R. (2005). Overview of the CLEF 2005 Multilingual Question Answering Track. Working Notes of the CLEF 2005. Vienna, Austria, 2005.
11. Vicedo J. L., Rodríguez H., Peñas A., and Massot M. (2003). Los sistemas de Búsqueda de Respuestas desde una perspectiva actual. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural. Num.31, 2003.