A Shallow Approach for Answer Selection based on Dependency Trees and Term Density

Manuel Pérez-Coutiño[1], Manuel Montes-y-Gómez[2], Aurelio López-López[2]
Luis Villaseñor-Pineda[2] and Aarón Pancardo-Rodríguez[1]

[1]Vanguard Engineering Puebla (VEng)
Recta a Cholula 308-4, CP 72810, San Andrés Cholula, Pue., México
[1]{mapco, apancardo}@inaoep.mx

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

Abstract. This paper describes the experiments performed for the QA@CLEF-2006 within the joint participation of the eLing Division at VEng and the Language Technologies Laboratory at INAOE. This year our laboratories have participated in the Spanish monolingual task, continue with their previous work described in [Pérez-Coutiño et al., 2005]. The aim of these experiments was to observe and quantify the possible improvement at the final step of the Question Answering prototype when some syntactic features were taken into the decision process. In order to reach this goal, a shallow approach to answer ranking based on the term density measure has been included. This measure weighs the number of question terms which have a syntactic dependency to one candidate answer within a relevant passage to the given question. Once the term density has been computed for each candidate answer, their weights along to the weights gathered in the previous steps are merged by a lineal combination to gather the final weight for each candidate answer. Finally, the answer selection process arranges candidate answers based on their weights, selecting the top-*n* as the Question Answering system answers. The approach described has shown a small but interesting improvement against the same Question Answering prototype without this module. Nevertheless, there are many variables to consider for a substantial improvement of the whole Question Answering system, and particularly at the initial steps, where passage retrieval and candidate answer selection are determinant for the improvement of system's recall.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [Database Managment]: Languages—Query Languages

General Terms

Measurement, Performance, Experimentation

Keywords

Question Answering for Spanish, Lexical-Syntactic Context, Natural Language Processing.

1 Introduction

Over the last years, the research of Question Answering (QA) systems for European languages has shown an incremental growing both in interest as well as in complexity. Particularly, QAs for Spanish has been formally evaluated within the CLEF initiative since 2003 [Magnini et al., 2003] when just one QA system for Spanish was proposed. In QA@CLEF-2004 [Magnini et al., 2004], there were a total number of five proposed QA systems for Spanish developed by four research groups from Spain and the one developed by INAOE (Mexico). Finally in QA@CLEF-2005 [Vallin et al., 2005], the total number of proposed QAs arose to seven, developed by six research groups from Spain, one developed by INAOE, and the joint participation between INAOE and the Polytechnic University of Valencia (Spain). Those QA systems have explored different approaches to cope with the whole QA problem, analyzing several methodologies from purely data-driven [Vicedo et al., 2003; Montes-y-Gómez et al., 2005] to in-depth natural language processing (NLP) [Ferrés et al., 2005]. Despite the methods used within some particular QA approach, the improvement of factoid questions resolution has been measured.

Starting on 2003, the QA research interests within the Language Technologies laboratory at INAOE have been directed in two main lines, on one hand, the development of QA systems for Spanish applying shallow NLP techniques in order to gather information models from which the system is able to extract answers to factoid questions. On the other hand, the development of full data-driven approaches to cope with the extraction of answers for both, factoid and definition questions. Those approaches have obtained encouraging results within CLEF campaigns, including the best overall accuracy for Spanish monolingual task at QA@CLEF-2005 [Perez-Coutiño et al., 2005].

This paper shows the prototype developed as a shared effort between the recently formed eLing Division at VEng[1] and the Language Technologies laboratory at INAOE. This approach continues with the previous work of the authors [Perez-Coutiño et al., 2005] to cope with factoid questions resolution. The aim of these experiments was to observe and quantify the possible improvement at the final step of a Question Answering prototype (i.e. at the answer selection step), as a consequence of introducing some syntactic features to the decision process. In order to reach this goal, the following key points have been included in the QA prototype: i) A syntactic parser based on dependency grammars for offline processing; ii) a shallow technique to weigh the number of question terms which have a syntactic dependency to one candidate answer within a relevant passage to the given question (aka *term density*); iii) a lineal combination for merging the term density of each candidate answer along to the weights gathered in the previous steps to obtain its final weight. Once all these values are computed, the answer selection process arranges candidate answers based on their weights, selecting the top-*n* as the Question Answering system answers. The approach described in this document has shown a small but interesting improvement up to 15% against the same Question Answering prototype without this module. Nevertheless, there are many variables to consider for a substantial improvement of the whole Question Answering system, and particularly at the initial steps, where passage retrieval and candidate answer selection are determinant for the improvement of system's recall.

The rest of the paper is organized as follows, section two summarizes the architecture of the prototype; section three exposes the details of the new elements within the prototype architecture; section four discusses training and official results achieved by the system at QA@CLEF-2006. Finally section five draws the preliminary conclusions of these experiments and discusses further work.

2 Prototype Architecture

As stated before, the system developed is based on the previous works of the authors [Pérez-Coutiño et al., 2005], where the most important modification relies on the inclusion of syntactic features to the decision process at the answer selection module. Figure 1 shows the main blocks of the system. It could be noticed that factoid and definition questions are handled independently. This report is focused in the factoid questions resolution process, while details of the processes involved in the creation and use of definition patterns aimed to answering definition questions can be found in [Denicia-Carral et al., 2006][2].

Factoid question treatment consists of the following steps: *question processing*, which includes the extraction of named entities and lexical context in the question, as well as question classification to define the semantic class of the answer expected to respond to a given question; *documents processing*, where the preprocessing of the supporting document collection is done in parallel by a *passage retrieval system (PRS)*and a shallow NLP, including the syntactic analysis of the document collection; *searching*, where a set of candidate answers is obtained from the *modeled* passages retrieved by the PRS; and finally *answer extraction*, where candidate answers are weighted and ranked in order to produce the final answer recommendation of the system. Next paragraphs summarize the initial steps of the prototype whilst section three and four discuss the new ones.

2.1 Question Processing. Our prototype implements this step following a straight forward approach involving the next steps:

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

The first step is responsible of identify the semantic class of the expected answer. In the experiments performed with the training data set, we found that when the number of classes was minimal (just 3 classes: date,

[1] The eLing Division at Vanguard Engineering is a private effort of recent foundation to develop business applications based on Language Technologies and Knowledge Management. In order to receive further information contact to Manuel Pérez-Coutiño (mapco AT v-eng.com)

[2] In order to avoid redundancy between reports this year we decided to leave the discussion of answering definition questions out of the scope of this report. The reader can find a discussion of the method evaluated for definition questions in this volume under the title "INAOE at CLEF 2006: Experiments in Spanish Question Answering".

quantity and proper noun) it was possible to achieve similar results in precision to those achieved when we use more than five classes, for instance person, organization, location, date, quantity and other. Steps 2 and 3 produce information used later during the searching step, mainly to match questions and candidate answer context, contributing to the weighted schema.

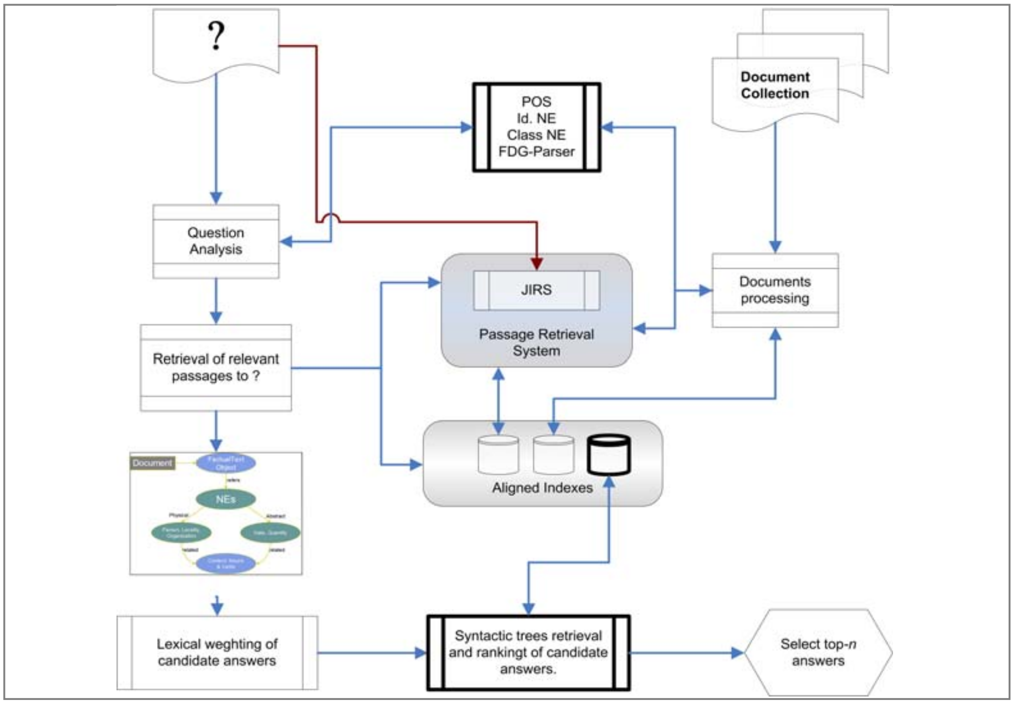


Figure 1. Block diagram of the system. Factoid and definition questions are treated independently. Factual questions require the following stages: question processing, documents processing, searching and answer selection. Notice the inclusion of a parser to the NLP tools and the resulting index tagged syntactically. Definition questions use a set of patterns for definition extraction and definition selection process.

2.2 Documents Processing. This year the prototype included an additional element for the syntactic parsing of the document collection. Thus the processing of target documents is composed of three parts, first the whole document collection is tagged with MACO[1], gathering the POS tags as well as named entities identification and classification for each document in the collection. In the second part of the process, and in parallel to the first one, the whole document collection is tagged with the FDG Parser from Conexor[3], which is based in the *Functional Dependency Grammar* discussed in [Järvinen & Tapanainen, 1997]. The final part of this step is performed by the JIRS [Gómez-Soriano et al., 2005] passage retrieval system (PRS), which create the index for the searching process. The index gathered by JIRS and the tagged collection are aligned phrase by phrase for each document in the collection. This way, the system could retrieve later the relevant passages for a given question with JIRS, and then use their tagged form for the answer extraction process.

2.3 Searching. This step is performed in two parts. The first part consists of the retrieving of relevant passages for the given question. This step is performed by JIRS, taking as input the question without previous processing. JIRS is a PSR specially suited for question answering systems. JIRS ranks the retrieved passages based on the computation of a weight for each passage. The weight of a passage is related to the lager n-gram structure of the question that can be found in the passage itself. The larger the n-gram structure, the greater the weight of the passage. A complete discussion of the similarity metrics used by JIRS and details of its evaluation can be found in [Gómez-Soriano et al., 2005].

Once the relevant passages are selected, the second part requires the POS and Parsing tagged forms of each passage in order to gather the representation used to extract candidates answers. Tagged passages are represented as described in [Pérez-Coutiño et al., 2004] where each retrieved passage is modeled by the system as a factual text object whose content refers to several named entities even when it is focused on a central topic. As mentioned, named entities could be one of these: persons, organizations, locations, dates, quantities and miscellane-

[3] http://www.conexor.com with an online demo of their software.

ous[4]. The model assumes that the named entities are strongly related to their lexical context, especially to nouns (subjects) and verbs (actions). Thus, a passage can be seen as a set of entities and their lexical context. Such representation is used later in order to match question's representation with the best set of candidates gathered from passages.

2.4 Answer Extraction. One of the drawbacks found in our previous work was the lost of precision during answer extraction step. Once the system applies different criteria to obtain a set of candidate answers, it computes for each candidate answers a lexical weight in order to rank the best candidate answers (see formula 1) and then select the top-*n* as the system answers. However, there were situations where several candidate answers could have the same weight. In order to avoid such situations, we have included syntactic information to compute an additional weight which is later combined with the lexical weight to obtain the final one used to rearrange the candidate answers. Next section discusses this concept.

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

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

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

3 Adding Syntactic Features

In order to improve the precision at the answer selection step, we have experimented with the inclusion of a shallow approach based on the use of some syntactic features. The main characteristic of this approach relies on the kind of information used to compute an additional weight to rearrange the set of candidate answers previously selected by a lexical supported method. This approach is flexible given that one of the factors taken in mind was to realize the amount of errors that a parser could yield.

3.1 Key observations of the method. There are different approaches to the use of syntactic information within the QA task. Some of them try to found a direct similarity between question structure and those of the candidate answers. This kind of similarity is commonly supported by a set of syntactic patterns which are expected to be matched by the answers. Some systems applying this approach are described in [Bertagna et al., 2004; Roger et al., 2005; Aunimo & Kuuskoski, 2005]. Other proposals, like the one presented by Tanev [Tanev et al., 2005] apply transformations to the syntactic tree of the question in order to approximate it to the trees of the relevant passages, where it is supposed that an answer is present.

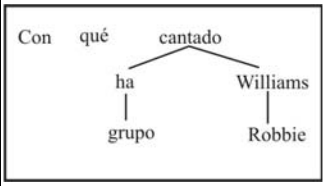
Despite the degree of effectiveness of those approaches, their principal drawback comes up when the parsing is deficient. With this limitation in mind, and trying to work around the gap of parsers for Spanish, our proposal is supported by the following observations of the dependency trees obtained by means of DFG parser.

1. There are important structural differences between questions dependency trees and those gathered from their relevant passages.
2. For a given question, its dependency tree states both, functional and structural dependencies starting from the main verb within the sentence, clearly delimiting other relations like subject, agent, object, etc. Contrasting this fact, dependency trees gathered from the relevant passage of the given question, could be seen as a forest, where functional and structural relations –which could possible lead the process to the accurate answer– are broken from one tree to another.

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

3. Finally, dependency trees gathered from the relevant passages to a given question could enclose a high number of question terms related to several candidate answers.

Figure 2 shows some examples of these observations.



Notice the differences between question and relevant passage structures. Besides, the analysis of the question shows the second case, whilst the passage tree could be seen as a forest. Finally, the tree where the answer is found *"Take That"* does not contain any question terms.

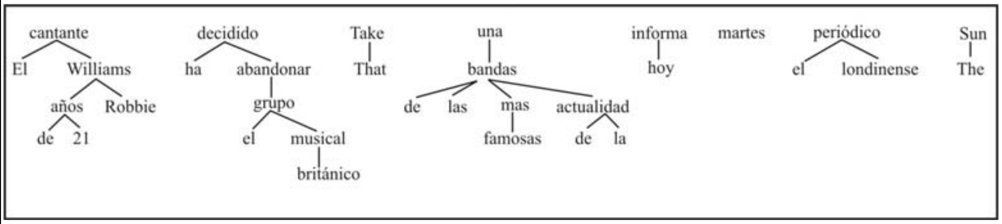
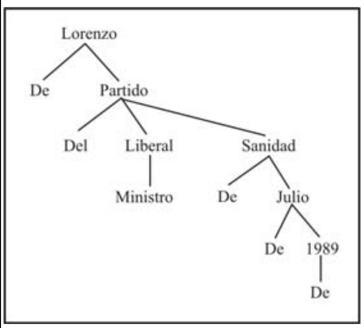


Figure 2a. Example of syntactic trees gathered from a question and one of its relevant passages. The answer is isolated from the rest of the terms within the dependency tree.



This example shows the tree gathered from the relevant passage to the question: *¿Qué político liberal fue ministro de Sanidad italiano entre 1989 y 1993?*

Notice that the accurate answer *"De Lorenzo"*, occurs within this tree, and relates several question terms to it.

Figure 2b. Example of one tree with several question terms related to its candidate answer (De Lorenzo).

3.2 Term Density. In order to cope with these observations, we propose a straight forward metric to capture and weigh up the question terms which are nearest to a candidate answer within a relevant passage. Formula 2 shows those relations, which we have named the *term density* within a dependency tree.

Given:

$Q = \{t_1, t_2 \ldots t_n\}$, the questions terms (lemmas)

$W = \{w_1, w_2 \ldots w_k\}$, the passages' terms (lemmas)

$C_i \subset W$, the terms (lemmas) of the *i-th* candidate answer.

A subtree $x_\alpha(w_1, \ldots, *, \ldots, w_h) : w_1 \ldots w_h$ depends on x_α (2)

Then, $\delta_q(C_i) = \frac{1}{n} \sum_{\forall t \in Q} f(t_j) \Leftrightarrow \forall w \in C_i, \langle w, x_\alpha \rangle$;

$$f(t_j) = \begin{cases} 1, \langle t_j, x_\alpha \rangle \text{ ó } t_j(*) \\ 0, \textit{else} \end{cases}$$

The algorithm applied in order to compute the term density for a candidate answer involves the following steps.

1. For each relevant passage of a given question
2. Retrieve the dependency tree for that passage
3. For each candidate answer within the passage
4. Retrieve the subtree where the candidate answer occurs
5. Apply formula 2 (δ_q)
6. Compute the maximum δ_q for each candidate answer, then preserve it if it is greater than 0.5, in other case, δ_q=0
7. Finally, computes the total weight of each candidate answer merging the lexical and syntactic weights with formula 3.

$$\omega(C_i) = \alpha\omega_{lex}(C_i) + \beta\delta_q(C_i) \quad (3)$$

Where ω_{lex} is the result of computing formula 1; α and β coefficients has been selected experimentally, giving them values $\alpha = 1/3$, and $\beta = 2/3$. This way the syntactic weight has a greater confidence.

4 Experiments and Results

This section discusses some results achieved with training questions as well as the experiment ran for the evaluation of the proposed prototype within the Spanish monolingual task at the QA@CLEF-2006.

4.1 Training Experiments

These experiments were performed over several training sets including the evaluation set of last year QA@CLEF. Through training experiments of the QA prototype we can observe a significant improvement in the final answer selection step. For the case of training with the QA@CLEF-2005 evaluation set, the accuracy of the system was increased over 7% for factoid questions, whilst the improvement in temporal restricted factoid questions achieved a 15%. These percentages represent a good progress giving that over the last years the rate of improvements in the results of Spanish factoid questions evaluation has been gradually.

Table 1 shows some examples of the increasing rank for candidate answers with high term density.

Table 1. Examples of rank increasing for candidates answers with high term density. Training set used was the evaluation set of the QA@CLEF-2005.

Question No.	Candidate and Right Answer	Previous Rank TO New Rank	$w_{lex}(C_i)$	$\delta_q(C_i)$	$w(C_i)$
29	De Lorenzo	5th TO 1st	0.5472	0.5000	0.5157
115	64 (días)	15th TO 1st	0.5440	0.5714	0.5622
139	Yoweri Kaguta Museveni	10th TO 1st	0.9367	0.8888	0.9048
161	Jacques Delors	3rd a 1st	0.8505	0.8000	0.8168

4.2 Evaluation

This year the evaluation of QA systems at CLEF included factoid, definition, temporal and list questions. The organizers provide to participants with a set of 200 unclassified questions, i.e. there were not markers to indicate the type of expected answer. Another novelty was that teams must provide answers with the specific passage where the answer was extracted from. The later was used to facilitate the evaluation of answers.

We participate in the evaluation with one run. The configuration applied considers three classes of possible answers, Date, Quantity and Proper Nouns; the system analyzes the first 100 1-line passages retrieved by JIRS; the lexical context used was formed with nouns, named entities, verbs, and adjectives, and the size of the window context is 8 words. Table 2 shows the results of the evaluation.

Despite the fact that our results (for factual questions) were only over 2% better than last year, we believe that the approach described could be a good starting point to the introduction of syntactic information to the answer selection process. Some errors observed while training include the confusion of accurate answers by candidate answers that have a term density similar or greater to the right answer. A detailed analysis of these results will help us to take the next direction in our research.

Table 2. Results of submitted run.

Run	*Vein061eses*
Right	84 (45F + 35D + 0 TRF)
Wrong	102
ineXact	3
Unsupported	5
Overall Accuracy	42.11%
Factoid Questions	30.82%
Definition Questions	83.33%
Temporally Restricted Factoid Questions	0%
Overall Confidence Weighted Score (CWS) over F, D, and TR	0.33582

5 Conclusions

This work has presented a method for the inclusion of syntactic information within the final process of answer selection in a QA system. The approach relies in the use of a flexible metric which allows measuring the amount of question terms which have a syntactic dependence to the candidate answers. Although official evaluation does not reach the expectation, preliminary results have demonstrated a significant improvement in the answer selection step. This drive us to think that it could be possible to apply syntactic information in several ways in order to cope with the problem of partial or even more, deficient syntactic trees (in particular dependence trees).

Despite the low increasing in the official evaluation, the methods applied at different steps of the QA process are stable; this conclusion can be dropped from the fact that the prototype has reached its last year performance.

It is important to realize that the additions to our QA prototype presented in this document are limited by the previous processes. This means that the proposed method is not able to extract new candidate answers. Therefore our next steps into QA systems development must be done in the direction of improving systems recall.

Acknowledgements. This work was done under partial support of Vanguard Engineering Puebla SA de CV, CONACYT (Project Grants U39957-Y and 43990), SNI-Mexico, and the Human Language Technologies Laboratory of INAOE. We also like to thanks to the CLEF as well as EFE agency for the resources provided.

References

[Aunimo and Kuuskoski, 2005] Aunimo L., Kuuskoski R., *Question Answering using Semantic Annotation*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Bertagna et al., 2004] Bertagna F., Chiran L., Simi M., *QA at ILC-UniPi: Description of the Prototype*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England

[Carreras and Padró, 2002] Carreras, X. and Padró, L. *A Flexible Distributed Architecture for Natural Language Analyzers*. In Proceedings of the LREC'02, Las Palmas de Gran Canaria, Spain, 2002.

[Denicia-Carral et al., 2006] Denicia-Carral C., Montes-y-Gómez M., Villaseñor-Pineda L. and García-Hernández R., *A Text Mining Approach for Definition Question Answering*, to appear in Proceedings for the Fifth International Conference on Natural Language Processing (FinTal 2006), Turku, Finland, August 2006.

[Ferrés et al., 2005] Ferrés D., Kanaan S., González E., Ageno A., Rodríguez H., *The TALP-QA system for Spanish at CLEF 2005*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Gómez-Soriano et al., 2005] Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., Rosso, P. *A Passage Retrieval System for Multilingual Question Answering*, to appear in the 8th International Conference on Text, Speech and Dialog, TSD, Springer LNAI, 2005.

[Magnini et al., 2003] Magnini B., Romagnoli S., Vallin A., Herrera J., Peñas A., Peinado V., Verdejo F. and Rijke M., *The Multiple Language Question Answering Track at CLEF 2003.*, in CLEF-2003 Workshop Notes, August 2003, Trondheim, Norway.

[Magnini et al., 2004] Magnini B., Vallin A., Ayache C., Erbach G., Peñas A., Rijke M., Rocha P., Simov K., Sutcliffe R., *Overview of the CLEF 2004 Multilingual Question Answering Track*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Montes-y-Gómez et al., 2005] 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"*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Pérez-Coutiño et al., 2004] M. Pérez-Coutiño, T. Solorio, M. Montes-y-Gómez, A. López-López and L. Villaseñor-Pineda, *"Toward a Document Model for Question Answering Systems"*, Advances in Web Intelligence: proceedings / Second International Atlantic Web Intelligent Conference AWIC04 Cancun, Mexico, May, 2004. Jesus Favela et al. (Eds.) LNAI 3034, pp. 145-154, Springer-Verlag 2004.

[Pérez-Coutiño et al., 2005] Pérez-Coutiño M., Solorio T., Montes-y-Gómez M., López-López M. and Villaseñor-Pineda L., "Question Answering for Spanish Supported by Lexical Context Annotation", to appear in proceedings of the 5th Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Peters C, et al. (Eds.), September 2004, Bath, England, Springer-Verlag 2005.

[Roger et al., 2005] Roger S., Ferrández S., Ferrández A., Peral J., Llopis F., Aguilar A., and Tomás D., *AliQAn, Spanish QA System at CLEF-2005*. 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Tanev et al., 2005] 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*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Tapanainen & Järvinen] Tapanainen P., Järvinen T., *A Non-Projective Dependency Parser*, in Proceedings of the 5th Conference on Applied Natural Language Processing, 1997.

[Vallin et al., 2005] Vallin A., Magnini B., Giampiccolo D., Aunimo L., Ayache C., Osenova P., Peñas A., de Rijke M., Sacaleanu B., Santos D., Sutcliffe R., *Overview of the CLEF 2005 Multilingual Question Answering Track*., in Working Notes for the Cross Language Evaluation Forum Workshop (CLEF-2005), Carol Peters (Ed.), September 2005, Vienna, Austria, ISTI-CNR, Italy 2005.

[Vicedo et al., 2003] Vicedo, J.L., Izquierdo R., Llopis F. and Muñoz R., *Question Answering in Spanish*., CLEF-2003 Workshop Notes, August 2003, Trondheim, Norway.