TALP-QA System for Spanish at CLEF-2004: Structural and Hierarchical Relaxation over Semantic Constraints

Alicia Ageno[1], Daniel Ferrés[1], Edgar González[1], Samir Kanaan[1], Horacio Rodríguez[1], Mihai Surdeanu[1], and Jordi Turmo[1]

TALP Research Center, Universitat Politècnica de Catalunya,
Jordi Girona 1-3, 08043 Barcelona, Spain
{ageno,dferres,egonzalez,skanaan,horacio,surdeanu,turmo}@lsi.upc.es
http://www.lsi.upc.es/~nlp

Abstract. This paper describes TALP-QA, a multilingual open-domain Question Answering (QA) system that treats both factoid and definitional questions. The system is described and evaluated in the context of our participation in the CLEF-2004 Spanish Monolingual QA task.
Our approach to solve factoid questions is to build a semantic representation of the questions and the sentences in the passages retrieved for each question. A set of Semantic Constraints (SC) are extracted for each question. An answer extraction algorithm extracts and ranks sentences that satisfy the SCs of the question. If matches are not possible the algorithm relaxes the SCs structurally (removing constraints) and/or hierarchically (abstracting the constraints using a taxonomy).
Answers to definitional questions are generated by selecting the text fragment with more density of those terms more frequently related to the question target throughout the corpus.

1 Introduction

This paper describes TALP-QA, a multilingual open-domain Question Answering (QA) system developed at UPC for the past 2 years. The paper focuses on our participation in the CLEF-2004 evaluation. Our aim in developing TALP-QA has been building a system as much as possible language independent, where language dependent modules could be substituted for allowing the system to be applied to different languages. A first preliminary version of TALP-QA for English was used for participating in TREC-2003 QA track (see [7]). From this initial version, a new version for Spanish was built and has been used in CLEF-2004. An improved version, again for English, has been used in TREC-2004.

In this paper we present the overall architecture of TALP-QA, we describe briefly its main parts, focusing on the components that have been changed more from our initial prototype, and on the components involving processing of the Spanish language. We present as well an evaluation of the system presented in the CLEF-2004 evaluation for both factoid and definition questions.

2 System Description

2.1 Overview

The system architecture follows the most commonly used schema, splitting the process into three phases that are performed sequentially. QA components may contain iterative algorithms (e.g. Passage Retrieval) but no feedback is propagated to the previous modules. There are three main subsystems (as shown in Figure 1), one corresponding to each phase: Question Processing (QP), Passage Retrieval (PR) and Answer Extraction (AE).

These subsystems are described below, but first we will describe some pre-processing tasks that have been carried out on the document collection (the EFE corpus in this case). As pointed out above, our aim is to develop a language independent system. Language dependent components are included only within the Question Pre-processing and Passage Pre-processing components, and can be easily exchanged with components for other languages.

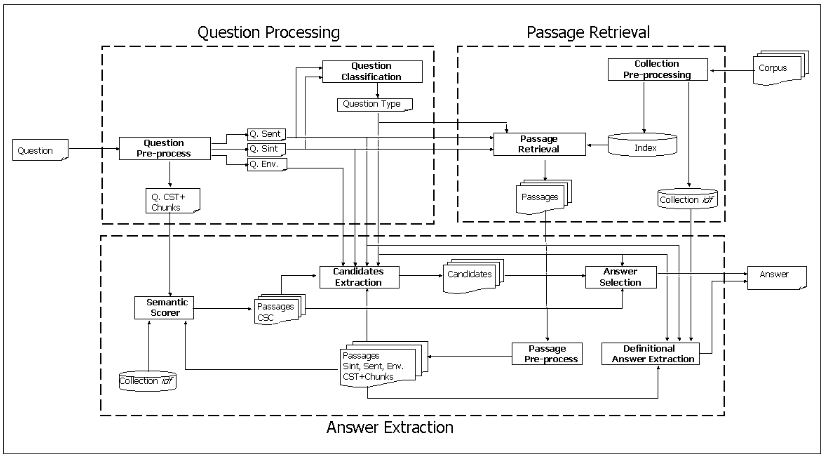

Fig. 1. Architecture of TALP-QA system.

2.2 Collection Pre-processing

In order to perform the PR task we have used the *Lucene*[1] Information Retrieval (IR) engine. Before CLEF-2004 we indexed the whole EFE collection: EFE 1994 and EFE 1995 (i.e. 454,045 documents). We pre-processed the whole collection with linguistic tools (described in next sub-section) to obtain the part-of-speech (POS) tags, lemmas and Named Entities (NE) of the text. This information has been used to built an index with the following parts:

[1] http://jakarta.apache.org/lucene

- Lemmatized and NE recognized text: this part is built using the lemmas of the words and the results of the NE recognition module. This part is indexed and used in the PR module.
- Original text with NE recognition: the original text that is retrieved when a query succeeds on the lemmatized text.

As an additional knowledge source that will be used in the AE task, an *idf* weight is computed at the document level on the whole colection.

2.3 Question Processing

The main goal of this subsystem is to detect the expected answer type and to generate the information needed for the other subsystems. For PR, the information needed is basically lexical (POS and lemmas) and syntactic, and for AE, lexical, syntactic and semantic. We use a language-independent formalism to represent all this information. In particular we use the same semantic primitives and relations for the two languages (English and Spanish) involved in our system.

For CLEF-2004 (for Spanish) we have used a set of general purpose tools of UPC NLP group (see [3] and [1]). The same tools are used for the linguistic processing of both the questions and the passages. These tools are:

- **FreeLing**, which performs tokenization, morphological analysis (including identification of quantities, dates, multiword terms, etc.), POS tagging and lemmatization.
- **Tacat**, a partial parser that obtains shallow nominal, prepositional and verbal phrases.
- **ABIONET**, a Named Entity Recognizer and Classifier that classifies NEs in basic categories (person, place, organization and other). See [2].
- **EuroWordNet (EWN)**, used to obtain the following semantic information: a list of synsets (with no attempt to Word Sense Disambiguation), a list of hypernyms of each synset (up to the top of each hypernymy chain), the EWN's Top Concept Ontology (TCO) class [10], and the Magnini's Domain Codes (DC) [5].
- **Three Gazetteers**, with the following information: acronyms, obtained using a Decision Tree approach [4], relations location-nationality (España-español, Spain-Spanish) and relations actor-action (escribir-escritor, write-writer).

The application of these linguistic resources and tools, language dependent, to the text of the question is represented in two structures (see Figure 2):

- **Sent**, which provides lexical information for each word: form, lemma, POS tag (an Eagles compliant rich tagset was used), semantic class of NE, list of EWN synsets and, finally, whenever possible the verbs associated to the actor and the relations between locations and their nationality.

- **Sint**, composed by two lists, one recording the syntactic constituent structure of the question (basically nominal, prepositional and verbal phrases) and the other collecting the information of dependencies and other relations between these components.

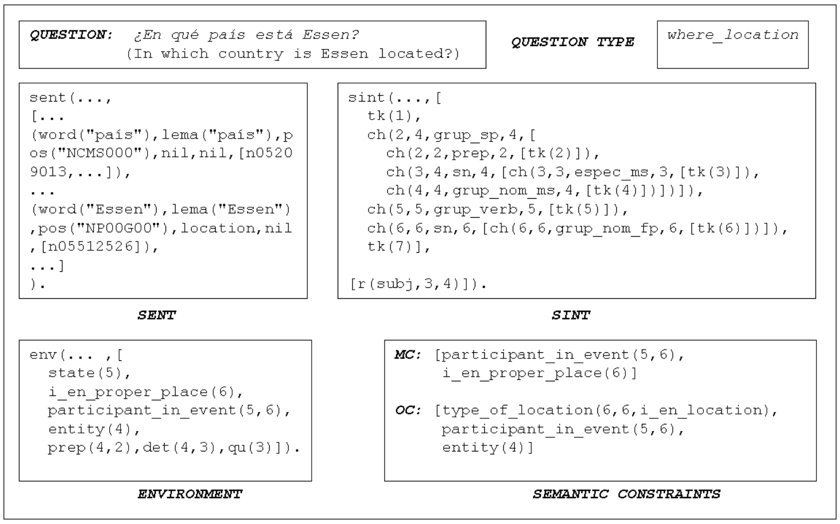

Fig. 2. Results of the pre-process of a question.

Once this information is obtained we can get the information relevant to the following tasks:

- **Question type**. The most important information we need to extract from the question text is the Question Type (QT), because all the work the system has to perform for searching the answer is based on this issue. A failure on identifying QT practically disables the correct extraction of the answer. Currently we are working with about 25 QT. The QT tries to focus the type of the expected answer and to provide additional constraints. For instance, when the expected type of the answer is a person, two types of questions are considered, *Who_action*, which indicates that we are looking for a person who performs a certain action and *Who_person_quality*, that indicates that we are looking for a person having the desired quality. The action and the quality are the parameter of the corresponding QT. The following are examples of questions correctly classified respectively as *Who_person_quality* and *Who_action* type:

  - *¿Quién fue jefe del XII Gobierno de Israel? (Who was the head of the XII Israel government?)*

- *¿Quién ganó el Premio Nobel de Literatura en 1994? (Who won the Nobel Prize in Literature in 1994?)*

In order to determine the QT our system uses an Inductive Logic Programming (ILP) learner that learns a set of weighted rules, from a set of positive and negative examples. We have used as learner the FOIL system [9]. A binary classifier (i.e. a set of rules) has been learned for each QT. As training set we have used the set of questions from TREC 8 and 9 (∼900 questions) manually tagged and as test set the 500 questions from TREC 11. All these questions were previously manually translated into Spanish. For each classifier we have used as negative examples the questions belonging to the other classes. For the classification task the following features have been used: Word form, Word position in the question, Lemma, POS, Semantic class of NE, Synsets together with all their hypernyms, TCO, DC and subject and object relations.
The set of rules for each class has been manually revised and completed with a set of manually built rules (with lower weights) in order to assure a greater coverage. See below a couple of such rules:

- A learned rule:

```
regla(non_human_actor_of_action,A,weight_1000,[],TT) :-
  sent(A,_,TT), TT=[_,W2|_],
  has_tco(W2,cObject),has_domain(W2,dTransport).
```

- A manual rule:

```
regla(non_human_actor_of_action,A,weight_994,[T1,T3],T) :-
  sent(A,_,[T1|T]), the_lema(T1,lema("qué")),
  has_chunk_with_hypernym(_,T,[T2|TT],
    [sArtifact,sObject,sAnimal],T3),
  the_pos(T2,pos("SP")),not(has_term_with_pos(TT,pos("AQ"),_)).
```

– **Environment.** The semantic process starts with the extraction of the semantic relations that hold between the different components identified in the question text. These relations are organized into an ontology of about 100 semantic classes and 25 relations (mostly binary) between them. Both classes and relations are related by taxonomic links. The ontology tries to reflect what is needed for an appropriate representation of the semantic environment of the question (and the expected answer). For instance, *Action* is a class and *Human_action* is another class related to *Action* by means of an *is_a* relation. In the same way, *Human* is a subclass of *Entity*. *Actor_of_action* is a binary relation (between a *Human_action* and a *Human*). When a question is classified as *Who_action* an instance of the class *Human_action* has to be located in the question text and its referent is stored. Later, in the AE phase, an instance of *Human_action* co-referring with the one previously stored has to be located in the selected passages and an instance of *Human* related to it by means of the *Actor_of_action* relation must be extracted as a candidate to be the answer.

The environment of the question is obtained from the *sint* and the information included in *sent*. To perform this task a set of about 150 rules has been built. The environment extracted from a question is presented in Figure 2.

– **Semantic Constraints.** The environment tries to represent the whole semantic content of the question. However, not all the items belonging to the environment are useful to extract the answer. So, depending on the QT, a subset of the environment has to be extracted. Sometimes additional relations, not present in the environment, are used and sometimes the relations extracted from the environment are extended, refined or modified. We define in this way the set of relations (the semantic constraints) that are supposed to be found in the answer. These relations are classified as mandatory, Mandatory Constraints (MC), (i.e. they have to be satisfied in the passage) or optional, Optional Constraints (OC), (if satisfied the score of the answer is higher). In order to build the semantic constraints for each question a set of rules (typically 1 or 2 for each type of question) has been manually built. A fragment of the rule applied in the example is presented in Figure 3. The rule can be paraphrased as follows: If the relation *state(C)* holds in the environment, get recursively all the predicates related to C, then filter the appropriate ones to be included in MC and OC and finally extend these sets for the sake of completeness. The application of the rule results in the constraints shown in Figure 2.

```
get_semantic_constraints(Question,MC,OC,Environment,where_location,1) :-
  …
  state(C,Question,Environment),
  get_related_tokens_in_environment(C,Environment,ListRelatedTokens),
  filter_tuple_tokens(ListRelatedTokens,MC,_,OC,
          [theme_of_event,time_of_event,location_of_event,which_entity],
          []),
  …
  filter_related_tokens(ListRelatedTokens,
          [
               [human_participant_in_event(C,_X)],
               [participant_in_event(C,_X), i_en_proper_person(_X)],
               [participant_in_event(C,_X), i_en_proper_organization(_X)],
               [participant_in_event(C,_X), i_en_proper_named_entity(_X)]
          ],
          MCRelations),
  …
  extend_mandatory(ListRelatedTokens,MCRelations,MC,OC,Question,Environment).
```

Fig. 3. A rule to obtain the Semantic constraints of a question.

2.4 Passage Retrieval

The main function of the passage retrieval component is to extract small text passages that are likely to contain the correct answer. Document retrieval is performed using the *Lucene* Information Retrieval system. For practical purposes we currently limit the number of documents retrieved for each query to 1000. The passage retrieval algorithm uses a data-driven query relaxation technique: if too few passages are retrieved, the query is relaxed first by increasing the accepted keyword proximity and then by discarding the keywords with the lowest priority. The reverse happens when too many passages are extracted. Each keyword is assigned a priority using a series of heuristics fairly similar to [8]. For example, a proper noun is assigned a priority higher than a common noun, the question focus word (e.g. "state" in the question "What state has the most Indians?") is assigned the lowest priority, and stop words are removed.

2.5 Factual Answer Extraction

After PR, for factual AE, two tasks are performed in sequence: Candidates Extraction (CE) and Answer Selection (AS). In the first component, all the candidate answers are extracted from the best scored sentences of the selected passages. In the second component the best answer is chosen.

- **Candidates Extraction**. This process is carried out on the set of passages obtained from the previous subsystem. These passages are segmented into sentences and each sentence is scored according to its semantic content regarding the question using the $tf * idf$ weighting of the terms from the question and their taxonomic neighbours occurring in the sentence (see [7]). The linguistic process of extraction is similar to the process carried out on questions and leads to the construction of the environment of each candidate sentence. The rest is a mapping between the semantic relations contained in this environment and the Semantic Constraints extracted from the question. The mandatory restrictions must be satisfied for the sentence to be taken into consideration; the satisfaction of the optional constraints simply increases the score of the candidate. The final extraction process is carried out on the sentences satisfying this filter.
  The Knowledge Source used for this process is a set of extraction rules owning a credibility score. Each QT has its own subset of extraction rules that leads to the selection of the answer. An example of extraction rule is presented in Figure 4. The rule can be paraphrased as follows: Look in MC for predicates *state(C)* and *location(X)* satisfied in the environment. Then look in the environment for the predicates, related to C, *location_of_event* and *location*. Make sure that the two locations are different and adjust the corresponding score.
  The process of application of the rules follows an iterative approach. In the first iteration all the semantic constraints have to be satisfied by at least one of the candidate sentences. If no sentence has satisfied the constraints, the set of semantic constraints is relaxed by means of structural or semantic relaxation rules, using the semantic ontology. Two kinds of relaxation are considered: i) moving some constraint from MC to OC and ii) relaxing some constraint in MC substituting it for another more general in the taxonomy.

If no candidate sentence occurs when all possible relaxations have been performed the question is assumed to have no answer.

```
extract_contextual_answer_from_tokens(DS,SS,_,_,Env, where_location,1, MT,A1,Sc2,_):-
  satisfy_MT_esp_obl([state(C),location(X)],MT,_),Sc=10,
  satisfy_strict([location_of_event(C,A,DS,Env),location(A,DS,Env)]),
  X\==A,
  nth(A,SS,A1),
  nth(X,SS,A2),
  smooth_scr(SS,X,A,Sc,Sc1),
if(
satisfy_MT_esp_obl([type_of_location(_,_,TL)],MT,_),
(check_type_of_location(A1,TL,A2,Sc3),Sc3 > 0.4, Sc2 is (Sc1 + Sc3 * 10)/ 2),
Sc2 is Sc1).
```

Fig. 4. One of the extraction rules applicable to the example.

– **Answer selection**. In order to select the answer from the set of candidates, the following scores are computed for each candidate sentence:

  - The rule score, which uses factors such as the confidence of the rule used, the relevance of the OC satisfied in the matching, and the similarity between NEs occurring in the candidate sentence and the question.
  - The passage score, that uses the relevance of the passage containing the candidate.
  - The semantic score, defined previously.
  - The relaxation score, which takes into account the level of rule relaxation in which the candidate has been extracted.

  For each candidate the values of these scores are normalized and accumulated in a global score. The answer to the question is the candidate with the best global score.

2.6 Definitional Answer Extraction

The approach taken to extract definitions can be viewed as a three-step process:

1. **Question analysis and target extraction**. The question is analyzed with the same module as factoid questions. This module outputs the target of the question and its type (human/organization). The type of the target allows to apply more specific heuristics to each question.

2. **Relative word significance computation**. The Relative Significance of a word stem is a measure of how related the word stem is to the question target; this relative significance is computed as follows. For each occurrence of the target in the corpus, a window with its 15 previous and following words is extracted. From each window extracted, adjectives and nouns (proper

and common nouns) are selected and stemmed (in order to reduce the high morphological variability of Spanish). This window is expected to capture the context of the target. Our observations recommend +/-15 words as an adequate distance, at least for Spanish.

The number of occurrences of each stem in the context windows is computed, and then multiplied by the *idf* of the stem as computed from the whole corpus, in order to obtain its relative significance to the target. Moreover, there are two lists of stems (one for persons and one for organizations) that contain stems likely to appear in definitions of either persons (as professions, prizes, etc.) or organizations (words like "partido", "organización"). The significance of stems appearing on the corresponding list (depending on the question target type) is multiplied by a factor determined experimentally (3.2) in order to boost its importance.

3. **Selection of the most informative fragment**. The definition has to be selected from the corpus. Definitions are usually found in fragments that follow some high-level patterns, as "<def> (<target>)" or "<target> , <def>". To obtain the definition, for each occurrence of one of these patterns in the text, what we call its information density is calculated, that is, the sum of the relative significance of its words divided by the number of nouns and adjectives it contains. The definition is expected to contain between 4 and 15 non-stop words, so the length of each definition is the one that maximizes its information density. The text fragment produced as final output is the definition with highest information density.

3 Results

This section evaluates the behaviour of our system in CLEF-2004. We evaluated the three main components of our system and the global results:

- **Question Processing**. This subsystem has been manually evaluated for factual questions (see Table 1) and the following components: basic NLP tools (POS, NER and NEC), semantic pre-processing (Environment, MC and OC construction) and finally, Question Classification. These results are accumulatives.
- **Passage Retrieval**. The evaluation of this subsystem was performed using the set of correct answers given by the CLEF organization (see Table 2). We submitted two runs. In both runs we retrieved only the 1000 top documents (no passages) for definitional questions. These runs differ only in the parameters of the passage retrieval module for factual questions:
  - Windows proximity: in run1 the proximity of the different windows that can compose a passage was lower than run2's (from 60 lemmas to 80).

Table 1. Results of Question Processing evaluation.

Subsystem	Total units	Correct	Incorrect	Accuracy	Error
POS-tagging	1667	1629	38	97.72%	2.28%
NE Recognition	183	175	8	95.63%	4.37%
NE Classification	183	137	46	74.86%	25.14%
Environment	180	81	99	45.00%	55.00%
MC	180	77	103	42.78%	57.22%
OC	180	131	49	72.78%	27.22%
Q. Classification	180	105	75	58.33%	41.67%

- Threshold for minimum passages: the PR algorithm relaxes the query to obtain more passages if the number of extracted passages is lower than this threshold. These values are: 4 (run1) and 1 (run2) passages.
- Number of passages retrieved: we have chosen a maximum of 3000 passages in run1 and 50 passages in run2.

Table 2. Passage Retrieval results.

Question type	Measure	run1	run2
FACTUAL	Accuracy (*answer*)	64.37% (103/160)	59.37% (95/160)
	Accuracy (*answer+docID*)	48.12% (77/160)	43.12% (69/160)
DEFINITIONAL	Accuracy (*answer*)	85.00% (17/20)	85.00% (17/20)
	Accuracy (*answer+docID*)	55.00% (11/20)	55.00% (11/20)

In this part we computed two measures: the first one (called *answer*) is the accuracy taking into account the questions that have a correct answer in its set of passages. The second one (called *answer+docID*) is the accuracy taking into account the questions that have a minimum of one passage with a correct answer and a correct document identifier in its set of passages.

- **Answer Extraction**. The evaluation of this subsystem for factual questions has been done in three parts: evaluation of the Candidates Extraction (CE) module, evaluation of the Answer Selection (AS) module and finally we performed an evaluation of the AE subsystem's global accuracy for factual questions in which the answer appears in our selected passages.
- **Global Results**. The overall results of our participation in TREC-2004 are listed in Table 4.

4 Evaluation and Conclusions

This paper summarizes our participation in the CLEF 2004 Spanish monolingual QA task. Out of 200 questions, our system provided the correct answer to 48

Table 3. Factual Answer Extraction results.

Subsystem	Measure	run1	run2
Candidates Selection	Accuracy (*answer*)	33.00% (34/103)	35.78% (34/95)
Answer Selection	Accuracy (*answer*)	70.58% (24/34)	79.41% (27/34)
Answer Extraction	Accuracy (*answer*)	23.30% (24/103)	28.42% (27/95)

Table 4. Results of TALP-QA system at CLEF-2004.

Measure	run1	run2
Total Num. Answers	200	200
Right/Wrong	48/150	52/143
IneXact/Unsupported	1/1	3/2
Overall accuracy	24.00% (48/200)	26.00% (52/200)
Accuracy over Factoid	18.89% (34/180)	21.11% (38/180)
Accuracy over Definition	70.00% (14/20)	70.00% (14/20)
Answer-string "NIL" returned correcty	19.23% (10/52)	20.37% (11/54)
Confidence-weighted Score	0.08780 (17.560/200)	0.10287 (20.574/200)

questions in run1 and 52 in run2. Hence, the global accuracy of our system was 24% and 26% for run1 and run2 respectively. We conclude with a summary of the system behaviour for the two question classes:

- **Factual questions**. The accuracy over factoid questions is 18.89% (run1) and 21.11% (run2). Although no direct comparison can be done with other evaluation in another language, we think that we have improved substantially our factual QA system with respect to the results of the TREC-2003 QA evaluation (5.3%) in English. In comparison with the other participants of the CLEF-2004 Spanish QA track (see [6]), our system has obtained the best results in the following type of questions: location, person and objects. On the other hand, our system has a poor performance in the classes: manner, measure, organization, other and time.
  - **Question Processing**. The Question Classification subsystem has an accuracy of 58%, a similar accuracy of the *environment*, MC and OC constraints. These values are influenced by the previous errors in POS, NER and NEC subsystems.
  - **Passage Retrieval**. In the PR we evaluated that 64.37% (run1) and 59.37% (run2) of questions have a correct answer in their passages. The evaluation taking into account the document identifiers shows that 48.12% (run1) and 43.12% (run2) of the questions are really supported.
  - **Answer Extraction**. The accuracy of the AE module for factual questions for which the answer occurred in our selected passages was of 23.32% (run1) and 28.42% (run2). It means that we achieved a significant improvement of our AE module, since the results of this part in TREC-2003 were 8.9%.

– **Definitional questions**. The definitional answer extraction module has obtained rather satisfactory results, 14 right definitions out of 20 proposed (70%), indeed the highest score for definitional questions in the Spanish language track. The errors are due to the shortage of passages retrieved for the target, which caused the module to fail to determine the right set of significant words.

Acknowledgments

This work has been partially supported by the European Comission (CHIL, IST-2004-506909) and the Spanish Research dept. (ALIADO, TIC2002-04447-C02). TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government.

References

1. Atserias, J., Carmona, J., Castellón, I., Cervell, S., Civit, M., Márquez, L., Martí, M.A., Padró ,L., Placer, R., Rodríguez, H., Taulé, M., Turmo, J.: Morphosyntactic Analisys and Parsing of Unrestricted Spanish Text. Proceedings of LREC-98. Granada, Spain.
2. Carreras, X., Márquez, L. and Padró, L.: Named Entity Extraction Using Adaboost. Proceedings of the CoNLL-2002. Shared Task Contribution. Taipei, Taiwan. September 2002.
3. Carreras, X., Chao, I., Padró, L., Padró, M.: FreeLing: An Open-Source Suite of Language Analyzers. Proceedings of LREC-2004. Lisbon, Portugal, 2004
4. Ferrés, D., Massot, M., Padró, M., Rodríguez, H., Turmo, J.: Automatic Building Gazetteers of Co-referring Named Entities. Proceedings of LREC-2004. Lisboa, Portugal, 2004.
5. Magnini, B., Cavagliá, G.: Integrating Subject Field Codes into WordNet. Proceedings LREC-2000. Athens, Greece, 2000.
6. Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the CLEF 2004 Multilingual Question Answering Track. Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004). Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany. 2005
7. Massot, M., Ferrés, D., Rodríguez, H.: QA UdG-UPC System at TREC-12. Proceedings of the TREC-2003. Gaithersburg, Maryland, United States, 2003.
8. Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Goodrum, R., Gîrju, R., Rus, V.: LASSO: A Tool for Surfing the Answer Net. Proceedings of the Text Retrieval Conference (TREC-8). Gaithersburg, Maryland, United States, 1999.
9. Quinlan, J.R.: FOIL: A midterm report. Proc. of the sixth European Conf. on Machine Learning. Springer-Verlag, 1993.
10. Rodríguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertanga, F., Roventini, A.: The Top-Down Strategy for Bulding EuroWordNet: Vocabulary Coverage, Base Conceps and Top Ontology. Computer and Humanities 32. 1998, Kluwer Academic Publishers.