The University of Évora's Participation in QA@CLEF-2007

José Saias and Paulo Quaresma

Departamento de Informática
Universidade de Évora, Portugal
{jsaias,pq}@di.uevora.pt

Abstract. The University of Évora participation in QA@CLEF-2007 was based on the Senso question answer system. This system uses an ontology with semantic information to support some operations. The full text collection is indexed and for each question a search is performed for documents that may have one answer. There is an ad-hoc module and a logic-programming based module that look for answers. The solution with the highest weight is then returned. The results indicate that the system is more suitable for the definition question type.

1 Introduction

This paper describes the use of Senso Question Answer System in the Portuguese monolingual Question Answering (QA) task of this year's edition of Cross Language Evaluation Forum (CLEF). After two previous participations in 2004 [1] and 2005 [2], the Informatics Department of the University of Évora developed and tested this new system, based on the authors' previous work [3] and [4].

Besides the usual newspapers collections from Público and Folha de São Paulo, the system had to consider also the Portuguese articles from Wikipedia. It uses an ontology as a knowedge base with semantic information usefull in several steps along the process.

The next section explains the system architecture. The methodology is described with examples in section 3. The evaluation of the obtained results is presented in section 4. Finally, some conclusions and future work are pointed out in section 5.

2 System Architecture

Senso Question Answer System has five major modules: *Libs*, *Query*, *Solver*, *Ontology* and *Web Interface*. Figure 1 represents the way they are connected.

The *Libs Module* contains collections of text documents. These collections (Público and Folha de São Paulo from years 1994 and 1995, plus the Wikipedia documents) are seen as libraries that contain information needed for question

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 316–323, 2008.

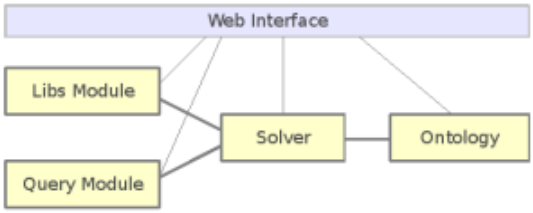

Fig. 1. Senso Modules

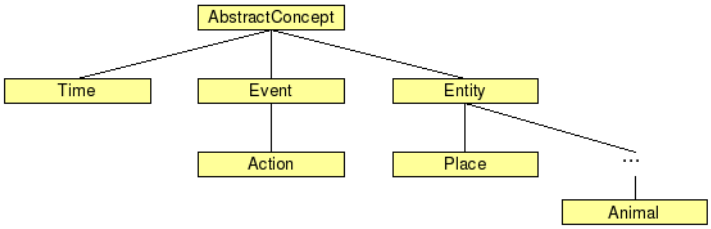

Fig. 2. Senso Ontology: top-level concepts

answering. All questions are firstly analyzed by the *Query Module*, which will select a set of relevant documents for each question, as explained later in section 3.

When we have an isolated sentence it's usually difficult to automatically capture its meaning. The *Senso Ontology* module has a starting knowledge base with semantic information that helps to perform the sentence analysis and the subsequent inference processes. This information is structured by an OWL[1] Ontology including concepts, relations and properties. Besides concepts and "IsA" relations, the ontology includes some simple facts about everyday life that might be very useful for text analysis. Our current ontology contains about 3500 concepts and has several relations connecting them: isA, usedFor, locatedAt, capableOf and madeOf. These concepts and relations represent a small common sense knowledge base about places, entities and events. Some of the top-level concepts are shown in figure 2.

The *Solver Module* performs a search for plausible answers in the identified relevant documents, being aware of the semantic expressed in the ontology. It has a logic-programming based tool and an ad-hoc answer selector.

The *Web Interface* layer allows an easier and friendly usage of the system, simplifying the analysis of each intermediate step in the process, as illustrated in figure 3. This interface is used to browse the ontology and make small changes to it, or to search for documents (or queries) and read them. Next section explains the methodology used to find the answers.

[1] OWL is the short name for Web Ontology Language. It is a language proposed by W3C to be used on *Semantic Web* for representation of ontologies.

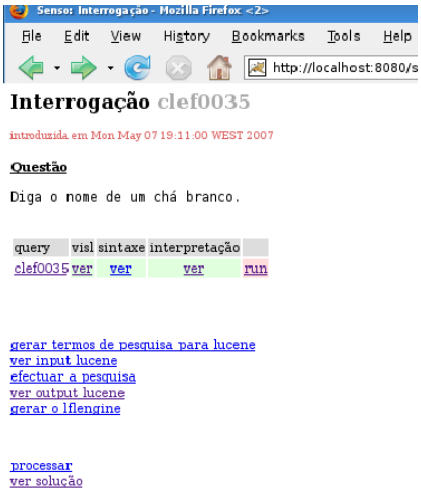

Fig. 3. Web Interface: options for intermediate analysis

3 Methodology

This section explains our approach to the Question Answer track in detail.

3.1 Import the Text Collections

The starting point is the information source: the document collections, having more than 500000 texts. The XML collection files were processed and split in single texts, along with their metadata. The *Libs Module* keeps all these individual documents, being aware of their temporal context, which is obtained from the collection.

Because we needed to perform some text search operations, the collections were indexed at this point with Lucene[2], a full-featured text search engine library. Each text was processed with PALAVRAS[5], a syntactical parser[3] based on the Constraint Grammars formalism that has a good coverage of the Portuguese language. This tool gives a detailed morpho-syntactical representation of the text for later usage.

3.2 Question Analysis

Each question is processed with the syntactical parser PALAVRAS[5] and a semantic analyzer able to obtain a partial semantic representation. The technique used for this process is based on Discourse Representation Structures (DRS) [7]. The partial semantic representation of a sentence is a DRS built with two lists,

[2] Apache Lucene is an open source project. http://lucene.apache.org/
[3] Tool developed by Eckhard Bick. VISL Project: http://visl.hum.sdu.dk/visl

one with the rewritten sentence and the other with the sentence discourse referents. We are only dealing with a restricted semantic analysis and we are not able to handle every aspect of the semantics. The DRS is a First-Order Logic expression which the logic resolution tool will try to understand.

Let us consider the following definition question, in this year's edition:

Quem é Boaventura Kloppenburg ? (Who is Boaventura Kloppenburg ?)

Figure 4 shows the morpho-syntactical representation given for that question. We can see the parser tags identifying the subject, the predicate and the interrogative form *quem* (*Who*). Figure 5 has the DRS for the same question, with the semantic representation used by the system for later logic inference process. That means the system will search for someone whose name is *Boaventura Kloppenburg*.

Our question answer system does a preliminary information retrieval task, in order to define a set of potentially relevant documents for each question. The amount of chosen documents may be from zero to several hundreds. This avoids the computational complexity of dealing with more than a half million texts. In the case where no candidate documents are found the system cannot find an answer and the result is NIL.

The *Query Module* creates the Lucene search query. This is done with the question text terms and, for some, their related terms. So, if a question has something like *"Which bird..."* the text search query will include synonyms of *bird* and specialization terms given by the Senso ontology, such as *eagle*. This semantic operation in the query allows the retrieval of a text that may not have the word *bird* but is still relevant as a possible answer source. As an example, one question asked which tree is present in the Lebanon flag. The answer was *cedro* (or cedar, in English). Being aware that cedar is a tree was important to the process.

```
QUE:fcl
=SC:pron-indp('quem' <interr> M/F S)    Quem
=P:v-fin('ser' PR 3S IND)       é
=SUBJ:prop('Boaventura_Kloppenburg' <org> F S)  Boaventura_Kloppenburg
=?
```

Fig. 4. Syntactical Parser: sample output

```
query(clef0012,
      [ name(_199, 'Boaventura_Kloppenburg' , ['F', 'S', 'Boaventura_Kloppenburg' ],
          [  ] ),
        q(_200, 'quem' , ['M/F', 'S', 'quem' ],
          [  ] ),
        'ser'(_199,_200,
          [ modif(verb,'ser', ['PR','3S','IND'] )  ] ),
        [ modif(ordObVSu), modif(casoSerNProp) ]   ],
      [ ref(_199), ref(_200) ] ).
```

Fig. 5. DRS for Question *'Who is Boaventura Kloppenburg ?'*

When the question belongs to a cluster and it is not the first from that group, the query is fed with more terms, in order to include the implicit topic. The system goes back to that cluster's first question and gets their search terms and answer into the Lucene query.

3.3 Solver Engine

The *Solver Module* is responsible for finding a list of answers for a query. Each answer has a weight and a snippet: sentence or expression justifying the answer and it's document identifier, as we can see in figure 6 for the question:

O que é um barrete frígio ? (What is a barrete frígio ?)

Fig. 6. Definition question result

The search for plausible answers is done on the Lucene selected documents by two tools: the *logic solver* and the *ad-hoc solver*. The semantic analyzer used before for the query will now create a DRS list for the selected texts. This list is a question dedicated Knowledge Base: the facts list. The *logic solver* is a logic-programming based module that performs a pragmatic interpretation of the query DRS over the full system knowledge base (the ontology and the facts list). It tries to find the best explanations for the question logic form to be true. This strategy for interpretation is known as "interpretation as abduction" [6].

The inference process is done with the Prolog resolution algorithm, which tries to unify the referents from the query with referents from documents, in the facts list, with help from the semantic information given by the ontology.

The *ad-hoc solver* is used for specific cases where a possible solution can be directly detected in the text. The system verifies each case conditions for the query and text expressions. Verifying the conditions might include a term semantic test for equivalence or "IsA" relation with another term, which is done by ontology analysis. Other conditions are related to text patterns, like *'X is DEFINITION'*, where the system attempts to learn the properties of *X*. This approach was used before in CLEF QA [8]. Figure 7 has a list of answers for the following question:

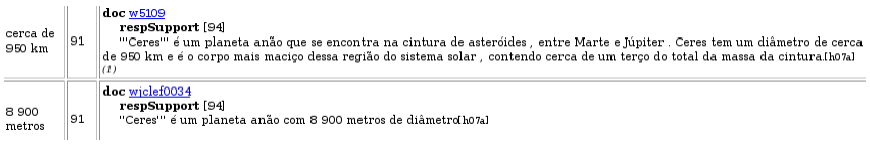

Fig. 7. Numerical factoid question result

Qual o diâmetro de Ceres ? (What is the diameter of Ceres ?)

This is a *Factoid* question about a measure. The ad-hoc solver identified the term *diâmetro* (diameter) and searched for numerical answers, including the unit of measure (*km, metros*).

There are cases where several documents lead the system to a common answer. This is the case in figure 8, where the *ad-hoc solver* found two documents with the same temporal expression as an answer candidate to a *When* question. This enforces that answer's weight.

The logic and ad-hoc found results are then merged to a final and weight sorted list. If the system finds more than one result for a question the QA@CLEF answer is the one with the maximum weight.

4 Results

In this QA@CLEF's edition, the Universidade de Évora's group registered for the monolingual Portuguese task, as did in previous participation [2], in 2005. A correct answer was found for 84 questions, which corresponds to an accuracy score of 42%.

Analyzing the results by question category, we can say that most of the errors were in the 90 wrong NIL returned values, where the system could not find an answer. Then, the *List* and *Temporally Restricted* questions represented a challenge and the obtained accuracy for these cases was around 20%. In the *Factoids* category the system had an accuracy close to the overall value, it was 39.62%. The best relative accuracy result was achieved in the *Definition* question type: 61.29%. Part of these definition answers were taken from Wikipedia documents, which sometimes had clear assertions. The overall Confidence Weighted Score over all assessed questions is 39.048/200 or 0.19524. All accuracy values for Portuguese as target are present in table 13 of QA Track Overview[9].

Comparing the current overall accuracy with the obtained in our department previous participation (25%) we believe this system produced good results. However, it needs some improvements as explained in the next section.

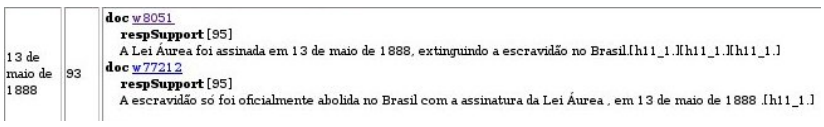

Fig. 8. Temporal Expressions

5 Conclusions and Future Work

In this paper we describe our Question Answering System for QA@CLEF-2007. Compared with the system we used in 2005, the Senso system has a different methodology and is based on a different ontology.

Analyzing the incorrect answers, we saw that some questions had no candidate documents where to search for an answer. This means that the Lucene query used for document retrieval failed in those cases. In other cases of wrong NIL answers, the system could not find a possible answer in the retrieved documents.

Our semantic analyzer also had some problems with DRS generation, while analyzing the morpho-syntactical representation of non-trivial sentences. Other problems were related to incorrect pragmatic analysis, in the logic solver, due to ontology limitations and some lack of precision on the semantic information taken from the text sentences.

The Lucene search engine indexes all text collections and gives a list of documents that may have an answer and need detailed analysis. This was important to avoid a problem we had in 2005, related to time constraints, because some of the hard work is now done only over the selected documents. We need to correct the way the Lucene text search query is built, to fetch the answer candidate documents where it currently cannot do it.

We also intend to improve the Senso ontology. Since many operations in our methodology depend on it's content, it should be manually revised and extended. Along with this, some disambiguation tool would help for better precision when a sentence concept is being related with an ontology existent term.

The *ad-hoc solver* is a rule based answer generator. This participation in CLEF shows that more rules are needed and some of the existing ones need an adjustment.

In a future participation, we intend to apply our system to other source languages, with Portuguese as target language. This might require an extra question translation module.

References

1. Quaresma, P., Quintano, L., Rodrigues, I., Saias, J., Salgueiro, P.: The University of Évora approach to QA@CLEF-2004. In: CLEF 2004 Working Notes (2004)
2. Quaresma, P., Rodrigues, I.: A Logic Programming Based Approach To QA@CLEF05 Track. In: CLEF 2005 Working Notes (2005)
3. Saias, J., Quaresma, P.: A proposal for an ontology supported news reader and question-answer system. In: Rezende., S.O., et al. (eds.) 2nd Workshop on Ontologies and their Applications (WONTO 2006) in the Proceedings of International Joint Conference, 10th IBERAMIA, ICMC-USP, Ribeirão Preto, Brazil (2006) ISBN: 85-87837-11-7
4. Saias, J., Quaresma, P.: A methodology to create ontology-based information retrieval systems. In: Pires, F.M., Abreu, S. (eds.) EPIA 2003. LNCS (LNAI), vol. 2902. Springer, Heidelberg (2003)
5. Bick, E.: The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press (2000)

6. Hobbs, J., Stickel, M., Appelt, D., Martin, P.: Interpretation as abduction. Technical Report SRI Technical Note 499, 333 Ravenswood Ave., Menlo Park, CA 94025 (1990)
7. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer, Dordrecht (1993)
8. Tanev, H.: Extraction of Definitions for Bulgarian. In: CLEF 2006 Working Notes (2006)
9. Giampiccolo, D., Forner, P., Peñas, A., Ayache, C., Cristea, D., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2007 Multilingual Question Answering Track. In: CLEF 2007 Working Notes (2007)