

The Senso Question Answering approach to Portuguese 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 team in QA@CLEF-2007 tested their SENSO system in the Portuguese monolingual task. The system uses an ontology semantic information for text search terms expansion and for verification of concept equivalency or *IsA*/specialization relations. The full text collection is indexed and for each question it's done a search, for retrieval of possible relevant documents that may have one answer. The solver module engine starts by producing a First-Order Logic expression representing the question and a logic facts list representing the texts information. There is a logic-programming based module that looks for answers within the facts list that unify and validate the question logic form. For cases where the answer can be directly detected in the text there is an ad-hoc module. Then the logic and ad-hoc found results are merged and the solution with highest weight is selected.

We sent one run result for evaluation. The overall accuracy was 42% and the Confidence Weighted Score was 0.19524. This paper has a description of our system.

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.3.7 Digital Libraries; I.2 [Artificial Intelligence]: I.2.3 Deduction and Theorem Proving—*Answer/reason extraction; Deduction; Logic programming*, I.2.7 Natural Language Processing

General Terms

Experimentation

Keywords

Natural Language Processing, Question answering, Questions beyond factoids

1 Introduction

This paper describes the use of SENSO Question Answer System in the Question Answering (QA) task of the 2007's edition of Cross Language Evaluation Forum (CLEF)¹. The presented system has been developed at the Informatics Department of the University of Évora, after two previous participations in 2004 [1] and 2005 [2].

The Portuguese monolingual QA had 200 questions that this system processed, looking for an

¹For more information, see <http://www.clef-campaign.org/> and <http://clef-qa.itc.it/2007/CLEF-2007.html>.

answer in the texts. Besides the usual newspapers collections from Público and Folha de São Paulo, the system had to consider also the Portuguese articles from Wikipedia.

Other relevant innovation was the existence of clusters or groups of questions about the same implicit topic, with possible anaphoric references between one question and the others. In such case, the system identifies the topic either in the first question or in the first answer, as referred in the guidelines.

This QA system (SENSO) is based on the authors previous work [3] and [4]. It uses an ontology as a knowledge base with semantic information usefull in several steps along the process.

The next section explains the system architecture. The followed 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.

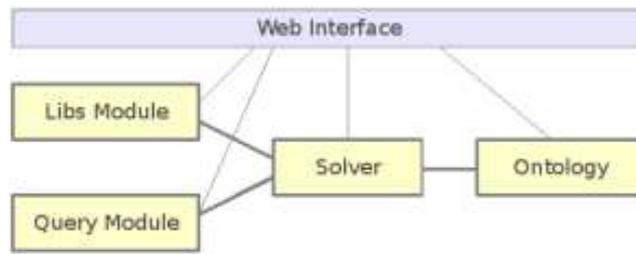


Figure 1: Senso Modules

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

All the questions are firstly analyzed by the *Query Module*. The query group identifier determines if a query will be associated with the first from that group. This module will also select a set of relevant documents for each question, as explained later.

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 OWL² Ontology including concepts, relations and properties. The OWL language has the intended semantic features and it is suitable for web publications, allowing us to share parts of our knowledge base in a direct and appropriate manner.

Besides concept "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.

²OWL[8] is the short name for Web Ontology Language and it is a language proposed by the W3C consortium to be used in the *Semantic Web* for the representation of ontologies. This language is based in the previous DAML+OIL (Darpa Agent Markup Language) language and it is defined using RDF (Resource Description Framework).

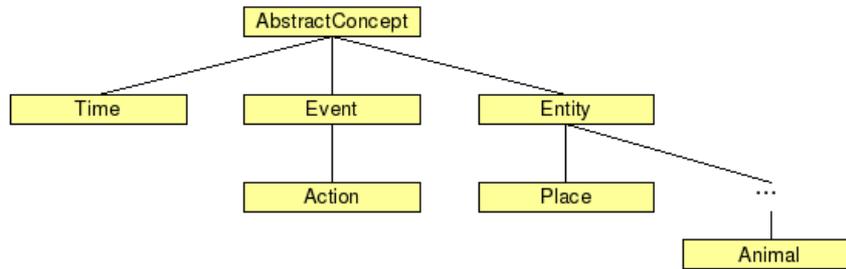


Figure 2: Senso Ontology: top-level concepts

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 to make small changes to it. We can also use a *web browser* application to search for documents (or queries) and read them.

Next section explains the methodology used to find the answers.

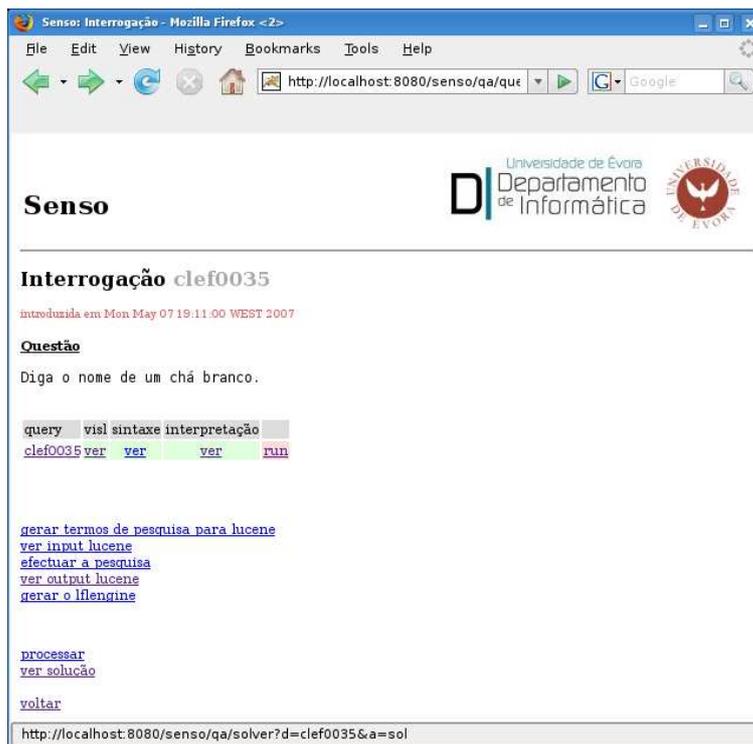


Figure 3: Web Interface: options for intermediate analysis

3 Methodology

This section explains our approach to CLEF-2007 Question Answer track in detail.

3.1 Import the Text Collections

The starting point is the information source: the document collections. This year we had 210734 newspaper texts and 336622 Wikipedia texts for the Portuguese target language. The XML collection files were processed and split in single texts, along with important 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 Lucene³, a full-featured text search engine library. Lucene scoring uses a combination of the Vector Space Model and the Boolean model to determine how relevant a given document is to a query.

Each text was then processed with PALAVRAS[5], a syntactical parser⁴ based in 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 latter 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, 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 ?

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 latter logic inference process.

```
QUE:fc1
=SC:pron-indp('quem' <interr> M/F S)    Quem
=P:v-fin('ser' PR 3S IND)                é
=SUBJ:prop('Boaventura_Kloppenbug' <org> F S) Boaventura_Kloppenbug
=?
```

Figure 4: Syntactical Parser: sample output

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

Figure 5: DRS for Question 12

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

³Apache Lucene is an open source project. <http://lucene.apache.org/>

⁴Tool developed by Eckhard Bick. VISL Project: <http://visl.hum.sdu.dk/visl>

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* produces 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. Question 15 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.

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 the core piece in our methodology. It is responsible for finding a list of answers for a query. Each answer has a weight and a support note: sentence or expression justifying the answer and it’s document identifier, as we can see in figure 6 for the question 17:

O que é um barrete frígio ?

Valor	Peso	Documentos que suportaram a resposta
uma espécie de touca ou carapuça	93	<p>doc w107696 respSupport [96] O "barrete frígio" ou "barrete da liberdade" é uma espécie de touca ou carapuça, originariamente utilizada pelos moradores da Frigia (antiga região da Ásia Menor, onde hoje está situada a Turquia).{h02}{h05}</p> <p>(1)</p>
barrete da liberdade	92	<p>doc w107696 respSupport [95] O "barrete frígio" ou "barrete da liberdade" é uma espécie de touca ou carapuça, originariamente utilizada pelos moradores da Frigia (antiga região da Ásia Menor, onde hoje está situada a Turquia).{h02ou}{h05ou}</p> <p>(1)</p>
utilizada pelos sincretistas helenistas e romanos	91	<p>doc w107696 respSupport [95] O barrete frígio é utilizada pelos sincretismo sincretistas helenistas e romanos, ainda que originalmente persa, deus salvador Mitra (divindade) Mithras .{h02}{h05}</p> <p>(1)</p>
familiar a alguns como o adorno de cabeça do Smurfs	91	<p>doc w107696 respSupport [95] O barrete frígio é familiar a alguns como o adorno de cabeça do Smurfs .{h02}{h05}</p> <p>(1)</p>

Figure 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 produce a DRS list for the selected texts. This list is seen as a small and 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 an answer generator for specific cases where the possible solution can be directly detected in the text. The system verifies each case specific conditions for the query and text expressions. When the conditions are verified, that ad-hoc case gives one answer. Verifying the conditions might include a term semantic test for equivalence or “IsA” relation with another term, which is done by ontology analysis.

Figure 7 has a list of answers for question 34:

Qual o diâmetro de 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*).

Valor	Peso	Documentos que suportaram a resposta
cerca de 950 km	91	doc w5109 respSupport [94] "Ceres" é um planeta anão que se encontra na cintura de asteróides , entre Marte e Júpiter . Ceres tem um diâmetro de cerca de 950 km e é o corpo mais maciço dessa região do sistema solar , contendo cerca de um terço do total da massa da cintura.h07a1 (1)
8 900 metros	91	doc wiclef0034 respSupport [94] "Ceres" é um planeta anão com 8 900 metros de diâmetro.h07a1 (1)
480 km	91	doc w5109 respSupport [94] No último dia de 1801, von Zach e Heinrich Olbers confirmaram a recuperação de Ceres.Ceres foi considerado demasiado pequeno para ser um verdadeiro planeta e as primeiras medidas apresentavam um diâmetro de 480 km.h07a1 (1)
250 km	91	doc w5109 respSupport [94] Existe alguma ambiguidade relativamente às características da superfície de Ceres. As imagens ultravioleta de baixa resolução tiradas pelo Telescópio espacial Hubble em 1995 mostram um ponto negro na sua superfície, ao qual foi dado o apelido Piazzi , que teria 250 km de diâmetro, um quarto da dimensão de Ceres, e que teria resultado do impacto de um asteroide com 25 km de diâmetro.h07a1 (1)
25 km	90	doc w5109 respSupport [94] Existe alguma ambiguidade relativamente às características da superfície de Ceres. As imagens ultravioleta de baixa resolução tiradas pelo Telescópio espacial Hubble em 1995 mostram um ponto negro na sua superfície, ao qual foi dado o apelido Piazzi , que teria 250 km de diâmetro, um quarto da dimensão de Ceres, e que teria resultado do impacto de um asteroide com 25 km de diâmetro.h07a1 (1)

Figure 7: Numerical factoid question result

The logic and ad-hoc found results are then merged to a final and weight sorted list. The answer merging process checks for repeated values and joins their support data, assigning the highest weight to that result. When 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. The proposed system was applied to the set of 200 questions and one run output was sent for evaluation.

A correct answer was found for 84 questions, which corresponds to an accuracy score of 42%. This value represents a relative increment of 68% from our department last participation global accuracy (25%).

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%. Table 1 shows the accuracy values in more detail.

Question Type	#	Right	Wrong	Unsupported	Inexact	Accuracy
Nil	111 returned	12	99	0	0	10.81%
Temporally Restricted	19	4	15	0	0	21.05%
Definition	31	19	5	0	0	61.29%
Lists	10	2	8	0	0	20.00%
Factoids	159	63	90	1	1	39.62%
All Questions	200	84	103	1	1	42.00%

Table 1: System accuracy results in QA@CLEF-2007

The overall Confidence Weighted Score over all assessed questions is 39.048/200 or 0.19524. Comparing the current overall accuracy with the obtained in our department previous participation we believe this system produced good results. However, it needs some improvements as explained in the next section.

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 in a different ontology. The results obtained are quite satisfactory and better since our last participation.

A preliminary analysis on the incorrect answers showed 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.

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 the system a list of documents that may have an answer and need detailed analysis. This was important to avoid problems with time constraints, because some of the hard work is now done only over the selected documents. However, 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.

In a future QA@CLEF participation we intend to apply our system to other languages besides the Portuguese monolingual task.

References

- [1] Paulo Quaresma, Luis Quintano, Irene Rodrigues, José Saias and Pedro Salgueiro. *The University of Évora approach to QA@CLEF-2004*. CLEF 2004 Working Notes.

- [2] Paulo Quaresma and Irene Rodrigues. *A Logic Programming Based Approach To QA@CLEF05 Track*. CLEF 2005 Working Notes.
- [3] José Saias and Paulo Quaresma. *A proposal for an ontology supported news reader and question-answer system*. Solange Oliveira Rezende et al. (Eds): 2nd Workshop on Ontologies and their Applications (WONTO'06) in the Proceedings of International Joint Conference, 10th IBERAMIA, ICMC-USP, Ribeirão Preto, Brazil, 2006. ISBN: 85-87837-11-7.
- [4] José Saias and Paulo Quaresma. *A methodology to create ontology-based information retrieval systems*. Fernando Moura Pires and Salvador Abreu (Eds): Progress in Artificial Intelligence - Proceedings of the 11th Portuguese Conference on Artificial Intelligence, EPIA'03, Beja, Portugal, 2003. Springer-Verlag, ISBN: 3-540-20589-6
- [5] Eckhard Bick. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.
- [6] Jerry Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. *Interpretation as abduction*. Technical Report SRI Technical Note 499, 333 Ravenswood Ave., Menlo Park, CA 94025, 1990.
- [7] Kamp, H. and Reyle, U. *From Discourse to Logic*. Kluwer: Dordrecht. 1993
- [8] Michael Smith, Chris Welty and Deborah McGuinness. *Owl web ontology language guide*. Technical report, 2004. <http://www.w3.org/TR/owl-guide/>