A proposal for a Web Information Extraction and Question-Answer System

José Saias[1] and Paulo Quaresma[2]

Departamento de Informática
Universidade de Évora, Portugal
jsaias[1]|pq[2]@di.uevora.pt

Summary. The Web is part of today's life and offers all kind of content. We present a system that can help the user to extract information from web documents and to find the answer for simple questions in natural language. This work is focused on newspaper articles and it is based on an ontology knowledge representation, natural language processment and a logic-programming framework.

1 Introduction

In the last decade the volume of available information on the web has grown exponentially. As an effect of globalization, the news we hear from a remote point of the globe have now gained importance and may influence some aspects of our life. In the other hand, most of the information taken in media resources may not be relevant to the end citizen. Nowadays, the main newspapers have an online RSS[1] service where they publish the latest news to all Internet users. Computer based systems can help people, allowing a quick and broader analysis on the available sources. This paper proposes an ontology based methodology for news article processing[2] in order to cover a large amount of documents, try to automatically understand some information in those documents and get automatic answers to some simple questions.

2 Common sense Knowledge Base

When we have an isolated sentence it's usually difficult to automatically capture the semantics in it. Ontologies allow the definition of class hierarchies, object properties and relation rules, such as, transitivity or functionality. Our

[1] Really Simple Syndication (sometimes also used for Rich Site Summary), is a popular XML format for Web content publication.
[2] This paper is an extension of previous work described in [5].

approach uses an ontology as the starting knowledge base with semantic information that helps to perform the sentence analysis and the subsequent inferences and interrogations. The ontology is expressed in OWL[3]. This 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 the formal concept definitions and "IsA" relations, there are a few simple facts about everyday life that might be very useful for document analysis. Some of those are also expressed by ontology relations. 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: *AbstractConcept* (the root concept), *Event*, *Time* and *Entity*. The next section explains the document analysis performed by the system.

3 Fetching and processing the news

Some popular newspapers like *Público* or *Correio da Manhã* have a "last hour" news section in their web site [4] , including an RSS channel. This is suitable for an automatic search for any recently added news article.

We used a program to periodically collect the recent news from *Público*'s RSS channel. As we can see in figure 1, each news item has some metadata fields: title, description, author, category, publication date and hour, and of course, the link to the web document containing the information. The category gives us a first simple classification for the document, placing it in Economy, Politics, International or Sports (in Portuguese Desporto - like the item listed in figure 1). The publication date gives the temporal context to the semantic content we find in the document, as we will see later. Each document imported to the system has a text body. That text is processed, following a methodology based on natural language processing techniques, namely, a syntactical parser and a semantic analyzer able to obtain a partial interpretation of the document. The tool used for the syntactical analysis is PALAVRAS [1]. It's a syntactical parser based in the Constraint Grammars formalism and it is able to cover a large percentage of the Portuguese language.
Let us consider a sentence in the above sports news item:

"Marcus Grönholm venceu neste domingo o Rali da Grécia." (in English: "Marcus Grönholm won the Greece Rally, this Sunday.")

The parser identifies the subject, the predicate and direct object with extra details that are stored on a Prolog structure and passed to the semantic analysis module. The technique used for this module is based on Discourse

[3] OWL is the short name for Web Ontology Language. It's a language proposed by the W3C Consortium for the *Semantic Web* and ontology representation.
[4] http://www.publico.pt/ and http://www.correiodamanha.pt

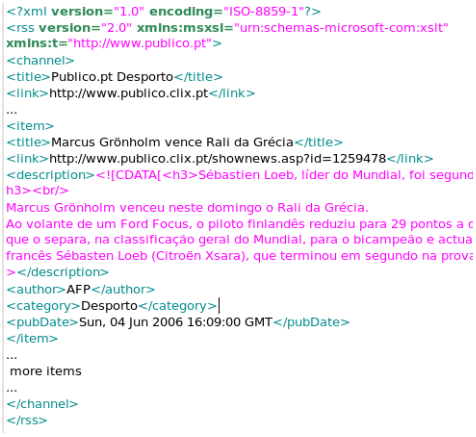

Fig. 1. RSS document from *Público*

```
item(publico1259478,
     'Desporto',
     'Sun, 04 Jun 2006 16:09:00 GMT').
...
sentence(publico1259478,
      [ name(A, 'Marcus_Grönholm' ,
                ['M/F', 'S', 'Marcus_Grönholm' ],
           [ ] ),
        name(B, 'Rali_da_Grécia' ,
                   ['M', 'S', 'Rali_da_Grécia' ],
           [ ] ),
        'vencer'(A,B,
           [ modif(verb,'vencer',
                        ['PS','3S','IND'] ) ] ),
        [ modif(temp,'domingo', ['M','S'],
                [ modif(pronDet,'este',
                           ['M','S'], []) ] )  ]   ],
      [ ref(A), ref(B) ] ).

... (other sentences)
```

Fig. 2. An item captured semantics

Representation Structures (DRS) [2]. 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: our focus is on the representation of concepts (nouns and verbs) and the correct extraction of its properties (modifiers, agents, objects). The previous news item is stored in the system with the details on figure 2.

4 Using the system

Once the news documents are obtained and analyzed they become part of the second knowledge base: the facts knowledge base. The Question-Answer module receives a natural language written query, in Portuguese. The query is processed using the same natural language tools used for the news texts. The search for an answer is done by a logic-programming based module that performs a pragmatic interpretation of the query DRS over the full system knowledge base (the ontology and the news facts).

The inference process is done with the Prolog resolution algorithm, which tries to unify the referent from the query with facts extracted from the documents and expressed in DRS structures.

4.1 Who/What Questions

As an example, we could enter a query like:
"Quem ganhou o Rali da Grécia?" (in English: "Who won the Greece Rally?")

The DRS for such query is presented in figure 3. This logic structure is checked against each sentence DRS. The result displayed by the system web

```
query(q281,
    [ q(X, 'quem' , ['M/F', 'S', 'quem' ],
        [  ] ),
     name(Y, 'Rali_da_Grécia' ,
          ['M', 'S', 'Rali_da_Grécia' ], [ ]),
     'ganhar'(X, Y,
         [ modif(verb,'ganhar',
             ['PS','3S','IND'] )  ] ),
     [  ]   ],
    [ ref(X), ref(Y) ] ).
```

Fig. 3. A *Who-Question* DRS

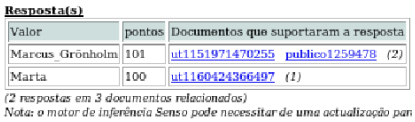

Fig. 4. Question-Answer result

interface is given in figure 4. It may include zero or more values considered valid as response to the query. For each possible response value there is also a document link list, pointing to the news item(s) where the system found the answer, and a numeric value with an estimated weight for that answer. Each sentence DRS component (subject, verb and object) match is given a weight (100 for direct match or less for dictionary and ontology driven cases). The weight for the document answer is calculated as the average weights of their matched sentence components. Finally, the weight assigned to an answer is the maximum value from their documents weights plus $\#docs - 1$. The previous question was answered because the concept *vencer* is defined as a synonym of *ganhar*. Another note is that there are two answers in the result. In this case, the reason is that we have a document with a sentence identifying last year winner. We could now follow the links and check the best solution by reading the text. The precise query for this year winner would be:
"Quem venceu o Rali da Grécia no ano de 2006?" (in English: "Who won the Greece Rally in the year 2006?")

That would introduce a temporal modifier on the query DRS expression to be checked against the date of publication of the document, such as:

```
... [ modif(temp,'ano', ['M','S'],
                [modif(num,'2006',
          [modif(prp,'de') ] )] )  ]   ], ...
```

The system answer is now only *Marcus*, as seen in the newspaper article.

4.2 When Questions

Another example of query about time is:
"Quando é que Marcus Grönholm venceu o Rali da Grécia?" (in English: "When did Marcus Grönholm won the Greece Rally?")

Once again, the interrogative term *quando*'s referent is matched against the temporal modifier on the sentence DRS: *"este domingo"* (in English: this Sunday). This information is then related with the news item publication date, the ontology and the sentence verb time (future, present or past), by the question-solver logic module. This allows the system to infer the desired date answer for the question, 2006-06-04. Similar treatment is given to temporal expressions like *today*, *this month*, *last year* and other. The next sentence list has several cases for temporal expressions:

```
- A Feira da Luz é em Setembro.
- A Feira da Luz é em Setembro de 1958.
- A Feira da Luz decorreu no último ano.
- A Feira da Luz foi no mês passado.
- A Feira da Luz é amanhã.
- A Feira da Luz é a 14 do próximo mês.
```

Each of these document sentences will produce an answer to the query: *"Quando é a Feira da Luz?"* (in English: "When takes place the Feira da Luz?")

The question-solver logic module infers the offset relative to the document publication date. Then presentation module gives a formatted date value. As an example, for the year 2006, if the document date is *Monday, October 2* and the sentence has *"... is on Thursday."* then the answer is next Thursday on that week: *2006-10-05*. If the document date is *Sunday, October 8* and the sentence has *"... will be in January."* then the answer is next January: *2007-01*.

4.3 Where Questions

For this kind of interrogations the sentence information near a preposition is taken into account and it is related with the ontology concepts below "*lugar*" (*place*, such as a *city* or *country*). If the term found in the selected sentence is a possible place, then it can be used as an answer. Lets consider the three documents in figure 5 and their assertions. Asking where is *Feira da Luz*, with a query: *"Onde é a Feira da Luz?"*
will give us the expected answers. Figure 6 has the result, one answer value per document and having equal weights. Those were direct answer cases. The system can also infer the answer for some nontrivial cases. Having the previous three assertions, we can ask if *Feira da Luz* is in a certain city:
"A Feira da Luz é em Montemor-o-Novo?"
and the answer is *yes*, because there is an indication, given by the ontology, stating that *Montemor* is an alias to *Montemor-o-Novo*. In the case where we ask if the event takes place in Portugal:
"A Feira da Luz é em Portugal?"
the answer is *yes*. This time it was not so immediate. The question-solver had to look for a place where *Feira da Luz* is happening and then check on the ontology or fact knowledge base if that place is located in Portugal.

5 Related Work

There are other initiatives related to the semantic content search. Ontologies are used in [3] for the specific domain of International Affairs. It has a natural

```
ut1160492122437 A Feira da Luz é no Alentejo.
ut1160492097626 A Feira da Luz é em Montemor.
ut1160492138407 A Feira da Luz é em Évora.
```

Fig. 5. Natural language assertions

Resposta(s)

Valor	pontos	Documentos que suportaram a resposta
Évora	100	ut1160492138407 *(1)*
Montemor	100	ut1160492097626 *(1)*
Alentejo	100	ut1160492122437 *(1)*

(3 respostas em 3 documentos relacionados)

Fig. 6. QA result: Where case

language interface also, but works with RDQL[5] instead of the Prolog logic resolution environment we adopted. The semantic archive features provided by [4] include means to annotate news materials and semantic search and browsing capabilities. This system runs inside the newspaper environment and uses a newspaper library specialized ontology, while the system we present works alone and outside the newspaper, allowing the use of many independent news sources, and our ontology is about common sense knowledge and not about a specific domain.

6 Conclusions and Future Work

We presented a web fact learning system focused on news texts from media sites. The system captures natural language texts, in Portuguese language, and performs information extraction for the question-answer feature. This feature is supported by the logic inference module, whose accuracy is affected by the quality of the ontology and the precision of the semantic information taken from the text sentences.
The ontology should be manually revised and extended. The semantic analysis can be improved if we add a tool to identify the inter-sentence anaphoric references. Along with this, some disambiguation tool is needed for better precision when a sentence concept is being related with an ontology existent term or when the system is trying to match a sentence with a query structure. Finally, the automatic question-answer system needs to be fully evaluated.

References

1. Eckhard Bick. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus University Press, 2000.
2. Kamp, H. and Reyle, U. *From Discourse to Logic.* Kluwer: Dordrecht. 1993
3. J. Contreras, V. Richard Benjamins, M. Blázquez, S. Losada, R. Salla, J. Sevilla, D. Navarro, J. Casillas, A. Mompó, D. Patón, Óscar Corcho, P. Tena, I. Martos. *A Semantic Portal for the International Affairs Sector.* In Proceedings of the EKAW 2004. pages 203-215. Springer, 2004
4. Pablo Castells, F. Perdrix, E. Pulido, M. Rico, V. Richard Benjamins, J. Contreras, J. Lorés. *Neptuno: Semantic Web Technologies for a Digital Newspaper Archive.* In 1st European Semantic Web Symposium, Greece, pages 445-458, 2004.
5. 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.

[5] RDQL is a query language for RDF based on SquishQL. For more detail visit http://jena.sourceforge.net/tutorial/RDQL/