Ontology supported news reader and question-answer system: a proposal

José Saias and Paulo Quaresma

Universidade de Évora, Portugal
jsaias|pq@di.uevora.pt

Abstract. Reading the news is a very time-consuming task. We present a methodology for a system that will automatically analyse the "last hour" news articles, offering semantic based features and real-time news reaction options.
This proposal 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. There are many more sources of information and each one seems to produce much more potentially relevant documents.
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. The newspapers, tv and other media spread the news from any event to the whole world.

The average citizen can read the papers and watch some news program on tv. However it's not possible to be aware of all occurrences in the world. In the other hand, most of the information taken in media resources may not be relevant to the end citizen.
All this has a special importance to professionals whose activity relies on news analysis, such as stock market brokers, military intelligence or economists.

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.
Filtering by date and section (politics, economy) is not enough for today's demands. There must be done some automatic work on the body text of each article in order to capture the expressed semantics in it. This might involve NLP

[1] Really Simple Syndication (sometimes also used for Rich Site Summary), is a popular XML format for Web content publication.

techniques and an inference enabled system. The semantic information captured from a document is stored in a knowledge base. Ontologies allow the definition of class hierarchies, object properties and relation rules, such as, transitivity or functionality. The information extracted from each news document will be given a formal representation, associated with an ontology. The resulting knowledge base has the facts list expressed by instances of ontology classes, in a semantic context. Then it will be possible to make some inferences about them.

This paper proposes an ontology based methodology for news article processing in order to:

- cover a large amount of documents
- identify the most relevant documents
- try to automatically understand some information in those documents
- set a notification or action to do in case of a certain 'thing' happens
- get automatic answers to some simple questions

The initial knowledge base is described in the next section. Section 3 shows the techniques used for news document analysis and knowledge base evolution. In section 4 we present system features and how they are accomplished. Finally, in section 5 some conclusions and future work are pointed out.

2 Common sense Knowledge Base

When we have an isolated sentence it's usually difficult to automatically capture the semantics in it. Our approach has a starting knowledge base with semantic information that will help to perform the sentence analysis and the subsequent inferences and interrogations.
Each element found on a sentence will be related to the starting knowledge base (we call it Senso), and also with the previous sentence semantics processed.

Senso is still in development and is the sum of a taxonomy of classes and a list of semantic information gathered from several ways.
Some influences were taken from previous works ([2] and [1]) where we had a top-level OWL ontology with some basic concepts. This ontology has a hierarchy of concepts used to organize the set of concepts mentioned in portuguese text documents.

OWL[4] 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).

We used OWL because it has de 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. This lead us to ConceptNet[3], which is a freely available common sense knowledge base and natural language processing toolkit. This tool gave us access to a semantic network presently available in two versions: concise (200,000 assertions) and full (1.6 million assertions) about spatial, physical, social, temporal, and psychological aspects of everyday life.

ConceptNet is available in english and our work needs portuguese language. We started by choosing a set of terms that we where interested on. Then we automatically followed their relations in the ConceptNet semantic network for a couple of nodes and collected those too. Having the terms and relations identified we run an automatic translation. Finally we filtered some wrong translations using another portuguese dictionary.

The top-level ontology and the ConceptNet translated terms were merged in a manual form. This work was done by several people using a web application where we could browse the database, insert new classes, update relations between classes, add or remove facts. Facts are also expressed in this resulting ontology, by class relations, class instances from an "action" concept or other semantic expression.

Figure 1 has a screenshot of the Senso KB Web Interface. We can see an search in the ontology for terms with a certain pattern (here: *ão*) and having an *IsA* relation with the term *animal*. The result of such query will include *dog* and *lion*, which in portuguese have the specified syntactic pattern, as shown on figure 2. Choosing any of those results, with a mouse click will show that term details, as illustrated on figure 3 for the term *firearm*. Those lines are part of that term details and they mean that *firearm* is a kind of weapon an might be used for actions like murder, hunt, shoot something or protect.

Senso knowledge base has relations connecting terms like *IsA*, *UsedFor*, *LocatedAt*, *CapableOf* and others.

Next section explains the document analysis performed by the system.

3 Fetching and processing the news

The proposed system might be used with any text documents in portuguese natural language. Is this case, we focus our work in news articles that are published day by day by the national media.

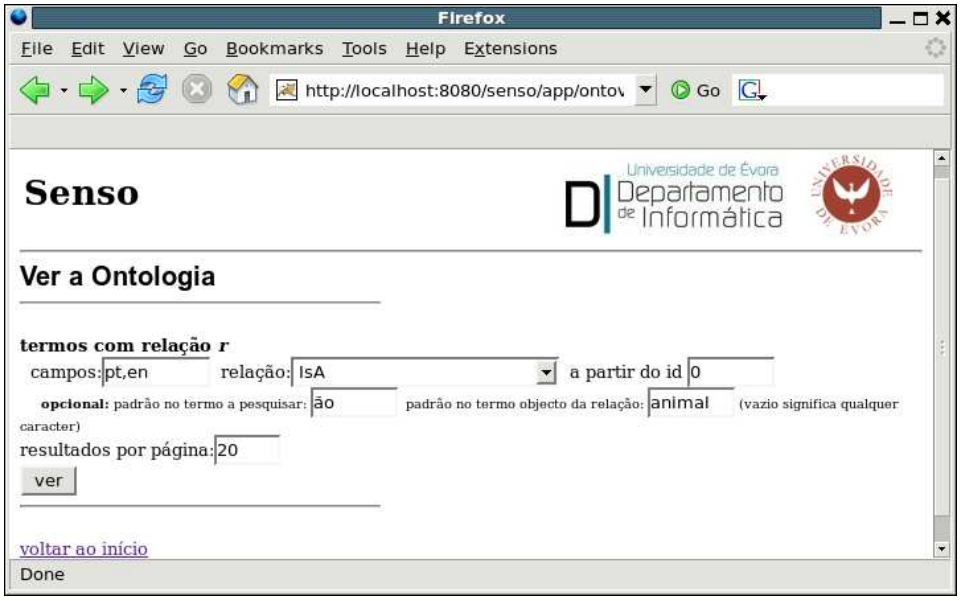

Fig. 1. Web interface for Senso Knowledge Base analysis

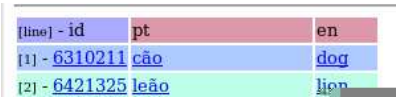

Fig. 2. Senso: partial query result

Some popular newspapers like *Público*[2] or *Correio da Manhã*[3] have a "last hour" news section in their web site, 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 4, 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 4). The publication date gives the temporal context to the semantic content we will find in the document. Later we will see some examples.

[2] http://www.publico.pt/
[3] http://www.correiodamanha.pt

arma de fogo	IsA	arma
arma de fogo	UsedFor	assassinato
arma de fogo	UsedFor	caça
arma de fogo	UsedFor	disparar
arma de fogo	UsedFor	proteger

Fig. 3. Senso: some details on *firearm* term

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rss version="2.0" xmlns:msxsl="urn:schemas-microsoft-com:xslt"
xmlns:t="http://www.publico.pt">
<channel>
<title>Publico.pt Desporto</title>
<link>http://www.publico.clix.pt</link>
...
<item>
<title>Marcus Grönholm vence Rali da Grécia</title>
<link>http://www.publico.clix.pt/shownews.asp?id=1259478</link>
<description><![CDATA[<h3>Sébastien Loeb, líder do Mundial, foi segundo</
h3><br/>
Marcus Grönholm venceu neste domingo o Rali da Grécia.
Ao volante de um Ford Focus, o piloto finlandês reduziu para 29 pontos a distância
que o separa, na classificação geral do Mundial, para o bicampeão e actual líder, o
francês Sébasten Loeb (Citroën Xsara), que terminou em segundo na prova grega.]]
></description>
<author>AFP</author>
<category>Desporto</category>
<pubDate>Sun, 04 Jun 2006 16:09:00 GMT</pubDate>
</item>
...
 more items
...
</channel>
</rss>
```

Fig. 4. Document source: RSS news channel from *Público*

Each document imported to the system has a text body. That text will be 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 [5]. It's a syntactical parser developed by E. Bick in the domain of the VISL Project[4]. This parser is based in the Constraint Grammars formalism and it is able to cover a large percentage of the Portuguese language.
Because the parser output is in a non-standard format, it was necessary to transform it into a structured form, like XML and Prolog terms. This was accomplished with the translation tool[5] Xtractor[6], that performs the conversion *VISL to Prolog and XML*.

[4] http://visl.hum.sdu.dk/visl
[5] It is also available to the other VISL users at http://abc.di.uevora.pt/xtractor

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.")

As can be seen in figure 5, the parser identified correctly the subject, the predicate and direct object.

Parser PALAVRAS output:

```
STA:fcl
=SUBJ:prop('Marcus_Grönholm' M/F S)     Marcus_Grönholm
=P:v-fin('vencer' PS 3S IND)    venceu
=ADVL:pp
==H:prp('em' <sam->)    em
==P<:np
===>N:pron-det('este' M S <dem> <-sam>) este
===H:n('domingo' M S <temp>)    domingo
=ACC:np
==>N:art('o' M S <artd>)        o
==H:prop('Rali_da_Grécia' M S)  Rali_da_Grécia
=.
```

And the Prolog representation:

```
sentence(syn(sta(fcl,
   subj(prop('Marcus_Grönholm','M/F','S'),'Marcus_Grönholm'),
   p(v_fin('vencer','PS','3S','IND'),'venceu'),
    advl(pp, h(prp('em','<sam->'),'em'),
      p(np, n(pron_det('este','M','S','<dem>','<-sam>'),'este'),
      h(n('domingo','M','S','<temp>'),'domingo'))),
    acc(np, n(art('o','M','S','<artd>'),'o'),
      h(prop('Rali_da_Grécia','M','S'),'Rali_da_Grécia', '.'))))).
```

Fig. 5. Syntactical analysis for a sentence

The next step is the semantic analysis. 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: our focus is on the representation on concepts (nouns and verbs) and the correct extraction of its properties (modifiers, agents, objects). In the last section of this paper we point out some possible improvements for this text semantic analysis.

The previous news item will be stored in the system with the following details:

```
item(publico1259478,
     'Desporto',
     'Sun, 04 Jun 2006 16:09:00 GMT').
...
sentence(publico1259478,
      [ name(A,'Marcus_Grönholm'),
        name(B,'Rali_da_Grécia'),
           vencer( A, B, [ modif( temp, ['este','domingo'] ) ]  )   ],
         [ ref(A), ref(B) ]  ).
```

... (other sentences)

Fig. 6. Representation for an item captured semantics

Notice that we store some metadata like the item category and the time of publication. This is sometimes useful, when a sentence has a temporal modifier that might be related to the publication field. Later we shall see it.

All documents retrieved from the newspaper RSS channel are processed using this technique. This produces a knowledge base that we will use for inference processes as we describe in the next section.

4 Using the system

Once the news documents are obtained from the source and analyzed they become part of the second knowledge base used by the system. This is called the facts knowledge base, and used with Senso allows the system to perform some inference and then offer interesting features.
Three main features are presented in the following subsections.

4.1 Using a better search filter for news

When we visit the newspaper site we can search for news items from category X or items where some pattern might be found in the text. This is not enough when we want more than syntactical search.
Suppose we want to find the items about rally drivers playing some instrument. This would be hard to do with simple syntactical searches.

Our system can receive a search instruction and retrieve the set of documents that validate that condition. The Prolog syntax for the above search condition would be:

```
searchItemList([ condition([], [piloto_de_rally(X), toca(X,Y), instrumento(Y)] ) ],
               L).
```

Suppose the document D1 has a sentence like:

"Marcus Grönholm tocou piano até aos 16 anos de idade."
(in english: "Marcus Grönholm played the piano until 16 years old.")

Our ontology in Senso tells us that *piano* is an *instrument*. Now, if there is a sentence in out fact knowledge base saying that *Marcus_Grönholm* is a *piloto_de_rally* (rally driver), then the document D will indeed be selected for the returned list of retrieved relevant documents, *L*.

This feature is also available in a web form in a high level manner. Here we present the logic version because it shows better the technical details involved. The prolog query *searchItemList/2* is processed in the system backend and the list L is then shown in an nice format to the user.

4.2 Trigger an action if "something happens"

Sometimes we want to schedule some action in the case of some certain event occurs. Our system allows users to set actions that will be triggered if a news item with a content that validates a defined condition is found.
At the moment, these actions actions are only e-mails with some message that will be sent to the user if any new item text validates a condition. As an example, suppose you are interested in economic transactions where a company buys another company (the portuguese word is empresa, bellow), but you only want to receive the alert message if the news occur in November. Then you can instruct the system with the following:

Action:

```
mail you@di.uevora.pt -s ‘‘ALERT: a Company is buying another company’’
```

Condition Syntax:

```
conditionList([ condition([ metadata(pubDate,after,’2006-11’) ],
                          [ empresa(X), comprar(X,Y), empresa(Y)] ) ]).
```

When a new item is fetched and processed, it's internal representation is checked for each user defined action condition list. The verified conditions will trigger the execution of the associated action.

4.3 Question-Answer

The most interesting feature of the system is the Question-Answer feature. This receives a natural language written query, in portuguese, and will search for answers, based on the collected news documents and the Senso ontology.

The analysis of a natural language query is split in three steps: Syntax, Semantics, and Pragmatics. Each query is processed using the same natural language tools used for the news texts. First, the received interrogation is parsed by the methodology described before. Like we did with the text sentences, each query syntactical structure is translated into a First- Order Logic expression (a DRS). The relevant difference is the special care to take with the interrogative term in the query.
Note that, at present, we are not able to deal with general unrestricted queries nor to translate them from a syntactical into a semantic structure. In fact this is a quite complex NLP problem and we have decided to deal only with specific subsets of the Portuguese language, namely, with interrogatives about specific domains.

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 (Senso ontology and facts from the news).
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.

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 would be:

```
query( q281,
     [ [quem-X,[]] ],
     [ ganhar(X, Y, [modif(verb,past)] ),
       nome(Y,'Rali_da_Grécia')           ] ).
```

And the result displayed by the system web interface is given is figure 7. Note that in this case there is only one result. The response given may include zero or more values considered valid as response. For each possible response there is also the link to the news item where the system found the answer.

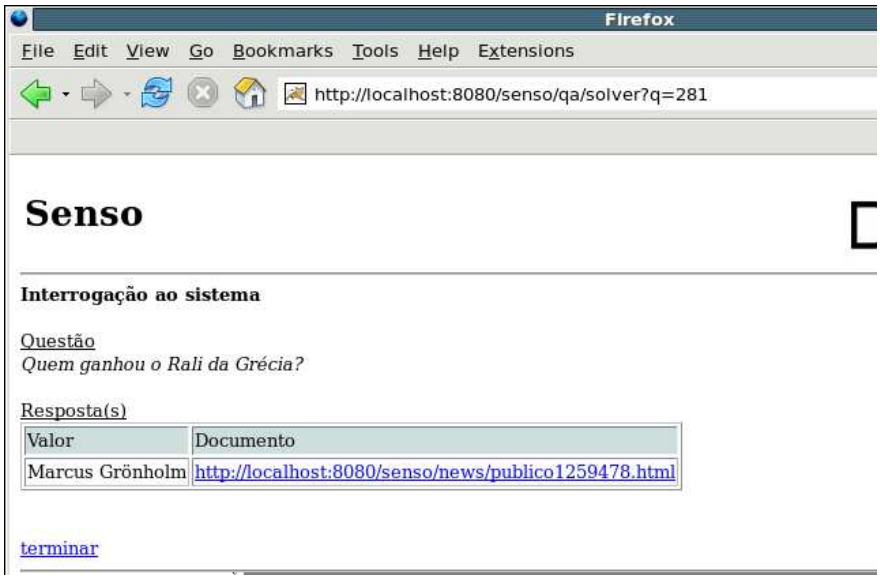

Fig. 7. Question-Answer result

The previous question was answered because the concept *vencer* is defined as a synonym of *ganhar*. Another note is that there was only one fact on the knowledge base about someone winning the Greece Rally. If we had a former document with a sentence identifying last year winner then probably we would get two answers. Then we could 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. 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 and we infer the desired date, 2006-06-04. Similar treatment is given to temporal expressions like *today*, *this month*, *last year* and other.

5 Conclusions and Future Work

We proposed a methodology to an ontology supported news reader and question-answer system. The system is based on the initial knowledge in the Senso ontology and on the semantic contents extracted from documents.
The system has described features which left us and some colleagues happy. However it exists only in a prototype version and it must now be improved to support a large-scale facts knowledge base.

The accuracy of the results found for automatic question-answer is also a point that needs more work. The system is affected in the inference process by:

- the quality of the Senso ontology
- 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. Finally, the system needs to be fully evaluated.

References

1. José Saias and Paulo Quaresma. *A methodology to create legal ontologies in a logic programming information retrieval system*, pages 185–200. V.R. Benjamins et al. (Eds): Law and the Semantic Web, LNAI 3369. Springer-Verlag, 2005. ISBN: 3-540-25063-8.

2. Paulo Quaresma, Luis Quintano, Irene Rodrigues, José Saias, and Pedro Salgueiro. *The university of evora approach to qa@clef-2004.* In Carol Peters, editor, *Question-Answering Track of the Cross Language Evaluation Forum 2004*, Bath, UK, September 2004.
3. H. Liu and P. Singh *ConceptNet: A Practical Commonsense Reasoning Toolkit.* BT Technology Journal, Volume 22, Number 4. Kluwer Academic Publishers, October 2004
4. Michael Smith, Chris Welty and Deborah McGuinness. *Owl web ontology language guide.* Technical report, 2004. http://www.w3.org/TR/owl-guide/
5. Eckhard Bick. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Aarhus University Press, 2000.
6. C. Gasperin, R. Vieira, R. Goulart and P. Quaresma. *Extracting XML syntactic chunks from portuguese corpora.* In TALN'2003 - Workshop on Natural Language Processing of Minority Languages and Small Languages of the Conference on "Traitement Automatique des Langues Naturelles", France, June 2003.
7. Kamp, H. and Reyle, U. *From Discourse to Logic.* Kluwer: Dordrecht. 1993