JaVaLI!: understanding real questions*

Luísa Coheur
L^{2}F INESC-ID/GRIL

Fernando Batista
L^{2}F INESC-ID/ISCTE

Joana Paulo
L^{2}F INESC-ID/IST

Spoken Languages Systems Laboratory
R. Alves Redol, 9, 1000-029 Lisbon, Portugal
{luisa.coheur, fernando.batista, joana.paulo}@l2f.inesc-id.pt

Abstract

This paper presents a linguistically motivated natural language processing system called JaVaLI![1], that transforms unrestricted text into logical forms.

Special focus is given to ambiguity and linguistic variation problems, which can be handled in different steps of the system processing chain.

Our system has been tested in the question interpretation domain and some preliminary results over a corpus of 680 questions in Portuguese are presented.

1 Introduction

Ambiguity contributes significantly to the complexity of linguistically motivated natural language (NL) systems. In fact, traditionally, if a word is ambiguous, the system has to deal with all possible values from the beginning of the processing chain, even if disambiguation only takes place some steps ahead.

Linguistic variations – the phenomenon in which similar semantic content may be expressed in different surface forms (Katz and Lin, 2000) – also increase the difficulty of translating text into logical forms, as different formulas are obtained for semantically equivalent sequences.

This paper describes JaVaLI!, a linguistically motivated natural language interface to a database of tourist resources, that translates unrestricted text into a formal language. As expected, the above mentioned problems – ambiguity and linguistic variations – contributed significantly to JaVaLI!'s complexity. In order to deal with these problems, special treatment of particular elements – such as interrogative pronouns and some special verbs – are applied in each step of the processing chain, simplifying our goal. In fact, authors such as (Katz et al., 1998) and (Milward, 1999) already highlighted the importance of handling structures – such as dates and compound nouns – in preprocessing modules, simplifying the processing chain.

This document is organized as follows: we start by presenting the different modules that compose JaVaLI!. A case-study about questions in the tourist resources domain is presented in section 3. Finally, a brief evaluation is conducted in section 4. The document ends with some concluding remarks and future work.

2 General architecture

JaVaLI! embodies several well compartmentalized sub-systems which, as Blitz's components (Katz et al., 1998), can be easily interchanged and switched on or off (Matos et al., 2003).

Morphossyntactic analysis is the first step performed using an external dictionary. The resulting text is then passed to a post-morphological processor that detects and forms special groups according to recomposition and correspondence

* This paper has been partially supported by FCT (Fundação para a Ciência e Tecnologia).

[1] JaVaLI! stands for "Já vamos na linguagem de interpretação!"

rules. This module groups the words into sentences and the resulting text is sent to a syntactic analyzer that, using a surface grammar, slips the sentences into nuclear phrases (structures from the chunk's family (Abney, 1995)). Chunks are then connected. Finally, and before creating the formulas in the representation language, the structures resulting from the previous step are transformed into a graph. The overall process (Figure 1) is described below in more detail.

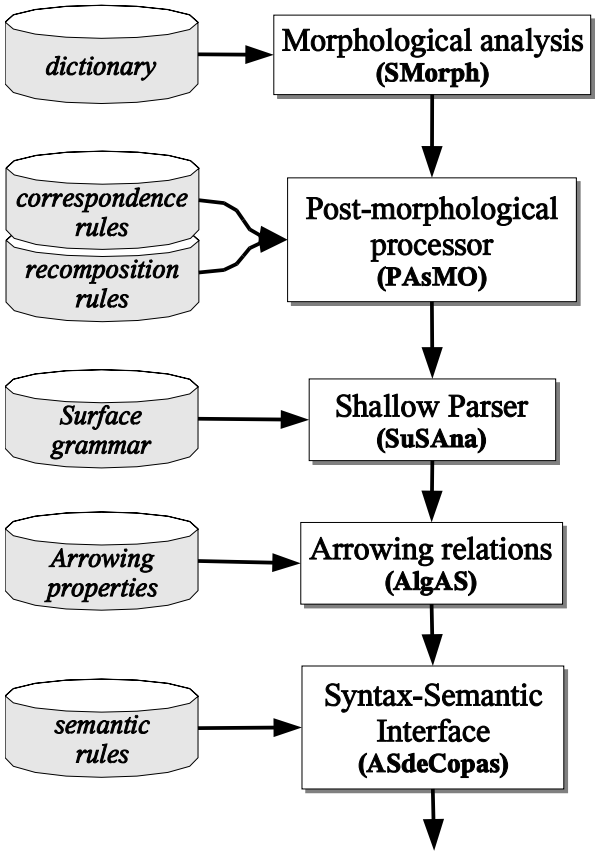

Figure 1: JaVaLI's architecture.

2.1 Morphossyntactic analysis

The morphossyntactic module enriches each word with its morphological characterization. For this task we use SMorph (Aït-Mokhtar, 1998) that also allows the construction of large dictionaries. The user declares the dictionary which is converted by SMorph into a compact binary file containing the correspondent finite state automata.

2.2 Post-Lemmatization Process

PAsMo (Paulo, 2001) is responsible for the Post-Lemmatization Process. Based on recomposition and correspondence rules, it processes sequences of words, such as dates, compound nouns, consecutive unknown words, numbers, and so on. It may also be used to translate tags, thus facilitating the interface with the syntactic analyzer. PAsMo was rebuilt from an old system (Faiza, 1999), improving its efficiency and expressivity.

2.3 Syntactic Analysis

The syntactic analysis is performed by SuSAna (Batista and Mamede, 2002), which produces a surface analysis where chucks are identified. Then, chunks are connected by arrows, which is done by AlgAS (Coheur, 2003). Finally, the resulting structure is converted into a graph.

Arrowing relations are somehow related with dependencies, but, contrary to the main dependency theories, arrows go from dependents to the head (Hagège, 2000). Moreover, the motivation behind an arrow relation is simply to connect two elements, because the established relations are needed to reach the desired semantic representation (details about this concept are given, *e.g.*, in (Bès, 1999; Bès and Hagège, November 2001)). To conclude, we state that each element is the source of at most one arrow and that no arrow's crossing is allowed.

2.4 Semantic Analysis

ASdeCopas[2] (Coheur and Mamede, 2003) is the module responsible for the syntax-semantic interface. Ideas, formalisms and data from the 5P paradigm (Bès, 1999; Bès and Hagège, November 2001; Hagège, 2000) are followed/used to reach ASdeCopas' input: a text with the associated graph. ASdeCopas uses hierarchically organized, intrinsically independent rules, that can be applied in any order, allowing information to be added incrementally. Also, partial results can be naturally produced. A formal presentation of the subsumption relation between rules is presented in (Coheur and Mamede, 2003).

[2] ASdeCopas stands for "Análise Semântica depois de Completada a análise sintáctica".

3 A practical application

This section starts by showing how to disambiguate the word *onde* (Portuguese word for *where*), which can be either a relative (onde_rel) or an interrogative adverb (onde_int).[3] This section also shows how to deal with linguistic variations. Finally, semantic rules dealing with interrogative elements are presented.

3.1 Dealing with ambiguity

Traditionally, morphological analysis shoots the system with both hypotheses, and posterior processing modules must deal with them, before disambiguation occurs. Instead, JaVaLI! – through SMorph – introduces a single ambiguous category – onde – covering both situations (dealing with ambiguous values is also suggested in (Poesio, 1994)). Then, during the processing chain, different levels of information are reached, allowing new disambiguation possibilities. This strategy prevents the complexity of dealing with (at least) two different categories from the beginning.

Notice that the disambiguation process cannot be done in a single magic moment: PAsMo resolves some ambiguity situations; then, after SuSAna's analysis, information about chunks is produced and other cases can be resolved; finally, AlgAS tries to disambiguate the remaining cases. At the end, reaching ASdeCopas, two things can be done to the remaining ambiguous values: either we treat them statistically, or we start to consider both categories.

The observation of the Edite corpus (see 4.1 for details) gave us the following heuristics to disambiguate the word *onde*:[4]

H1 if *onde* starts the question, it is a *onde_int*

ex: *Onde fica a serra mais alta de Portugal?*[5]

ex: *Onde se situam os parques naturais das Montanhas?*[6]

H2 if *onde* ends the question, it is a *onde_int*

ex: *O aqueduto das águas livres começa onde?*[7]

ex: *O museu de Arte Antiga é onde?* [8]

H3 if *onde* is followed by the sequence *é que*, it is a *onde_int*

ex: *Onde é que há coretos?*[9]

ex: *Onde é que se toca música folclórica?*[10]

H4 if another interrogative element – *quem*, *quais*, ... – was detected and *onde* is not coordinate with it, then it is a *onde_rel*

ex: *Qual é o maior Lago de Trás-os-Montes onde se pode andar de barco à vela?*[11]

ex: *Quais os lagos de Portugal onde posso praticar windsurf?*[12]

H5 if we have a exist-type question, it is a *onde_rel*

ex: *Existem campos de golfe onde possa ter lições em Alemão?*[13]

ex: *Existem parques de campismo no Gerês onde não seja necessário apresentar a carta de campista?*[14]

Finally, we present a last heuristic to disambiguate the word *onde*. It works fine in the observed corpus, but situations exist, where it won't apply. As so, this heuristic is presented as a last choice and it should be applied only if the previous heuristics had no success:

H6 if before the occurrence of *onde*, there are only prepositional phrases, it is a *onde_int*

ex: *Em Aveiro onde posso comprar peças de artesanato?* [15]

[3] In JaVaLI!, Categories are sets of attribute/value pairs, *i.e.*, feature structures hierarchically organized (see (Bès, 1999) for details). For exposure reasons we use a unique label to identify these sets.

[4] As we will see there is an order associated with these heuristics.

[5] Where is the highest mountain in Portugal?

[6] Where are the national parks in the mountains?

[7] Where does the águas livres aqueduct begins?

[8] The Museum of Ancient Art, is where?

[9] Where can one find a bandstand?

[10] Where is folkloric music played?

[11] Which is the biggest lake of Trás-os-Montes where one can sail?

[12] In which Portuguese lakes one can practice Windsurf?

[13] Is there any golf club where I can take lessons in German?

[14] Is there any camping park in Gerês where camping card is not necessary?

[15] Where can I buy handicrafts in Aveiro?

ex: *Em Setúbal onde posso jogar Andebol?* [16]

Following the process chain, we have that:

- PAsMo is able to apply H1, H2 and H3;
- After SuSAna's analysis, H4, H5 and H6 can be applied.

Disambiguation results are shown in section 4.

3.2 Linguistic variations

In this section we show how we deal with linguistic variations. Notice that we are not saying that we solve paraphrases. We limit ourselves in applying special treatments to particular sets of words, that we want to make converge to a certain formula.

Also, once again there is no gold moment to apply these treatments. For example, *porquê que* and *porque razão* are two ways of asking the same thing: why? PAsMo takes care of this constructions, grouping these elements together and transforming them into a single *porque*. However, this is not so simple: sometimes more or less complex sequences of words can occur between the elements that we want to group. This is the case of sequences as *Como se chama...*[17], where a prepositional phrase can occur between *como* and *se chama* (ex: *Como, em Lisboa, se chama..*)[18]. In these situations, only after detecting prepositional phrases – that is, after SuSAna's analysis – these elements can be grouped.

The arrowing concept also helps to handle linguistic variations as the same arrowing structures can capture several different syntactic structures. That is, different syntactic structures can be captured by the same arrowing relations, allowing to reach the same formulas. For example:

Onde fica o hotel Ritz?, *O hotel Ritz fica onde?*, *O hotel Ritz onde fica?* and *Fica onde o hotel Ritz*, they are all captured by the same arrowing relations.[19]

To conclude this section, the semantic rules itself can also ease the convergence process. For example, verbs *existir* and *haver*[20], in the forms *existe(m)* and *há*, respectively, may have different values.

- They may introduce an *exist-question*[21]
  - ex: *Há alojamentos no Gerês?*[22]
  - ex: *Existem praias com água quente na Costa Verde?*[23]
- They can mean *to have*
  - ex: *Em que praias do Alentejo há nadador salvador?*[24]
  - ex: *Em que praias do Alentejo existem nadadores salvadores?*[25]
- They can appear in contexts in which they don't have any relevant semantic value
  - ex: *Que casas do século 15 existem em Sintra?*[26]
  - ex: *Onde que há coretos?*[27]

By analyzing the involving context, we are able to decide which of the above situation we are dealing with. With this information we are able to converge *há* and *existe* into the same formal representation in the context of *exist-questions*. In addition, we can treat them as the word *com* (*with*) in the second situation. Finally, we can ignore them in the last case.

3.3 Semantic rules

In this section we briefly describe the semantic rules responsible for detecting the question type – where, when, what, *exist-question*... – and its target – an hotel, an event, ...

We start with a rule for the interrogative pronouns *que*[28], *qual*, *quais* and *quem*. They have in common the fact that they all arrow a name[29], and

[16] Where can I play handball in Setúbal?
[17] *What is the name....*
[18] ?*What in Lisbon is the name....*
[19] *Where is hotel Ritz?*
[20] Roughly, *to exist* and *to have*.
[21] *Does X exists?*
[22] *Is there any accommodation in Gers?*
[23] *In there any beach with warm water in Costa Verde?*
[24] *Which are the beaches in Alentejo having lifeguard?*
[25] *Which are the beaches in Alentejo having lifeguard?*
[26] *Which houses from the 15th century exist in Sintra?*
[27] *Where can one find bandstands?*
[28] It has a behavior similar of the *onde*.
[29] As *Quais são os hotéis ...* is equivalent to *Quais os hotéis ...*, we ignore the word *são*, and *Quais* arrows the noun *hotéis* in both cases.

they share the same semantic representation[30]. For example, both

... *qual a montanha*$_{166}$ ...

and

... *que montanha*$_{166}$...

translates into

?x$_{166}$, montanha(x$_{166}$)

In the same way, both

... *qual é o responsável*$_{656}$ ...

and

... *quem é o responsável*$_{656}$...

translates into

?x$_{656}$, responsável(x$_{656}$)

We also have a rule for onde_int, that can arrow either a noun or a verb:

Onde são os melhores hotéis$_{89}$ *do Algarve?*[31]
Onde é que se pode$_{124}$ *nadar na Costa Verde?*[32]

The obtained representation is

?local(x$_{89}$)

and

?local(x$_{124}$)

respectively. In the first situation we search the localization of the element identified by variable x$_{89}$, and in the second situation, a place where we can do something (in this case swimming).

In addition, we have rules for *quanto*[33], and for the family of the elements questioning time, such as *A que horas*, *Em que dias*, *Em que anos*.[34] Notice that these sequences where previously grouped by PAsMo.

Finally, we have rules that identify *exist-questions*, depending on the position occupied by the forms of the verbs *existir* and *haver*.

Before ending this section let us take a look at an example generated by JaVaLI!. Being given

Em que praias do Alentejo, com bandeira azul, há nadador salvador?[35]

the following formula is obtained:

```
R20:  ?x145
R1:   praias(x145)
R17:  de(x145, x148)
R23:  NAME(x148, Alentejo)
R17:  com(x145, x151)
R1:   bandeira(x151)
R9:   AM(x151, x152)
      id(x152)=azul
R35:  com(x145, x155)
R1:   nadador-salvador(x155)
```

Each R$_i$ identifies the applied rule. The predicate NAME associates a variable (representing an entity) with its name (as in (Allen, 1995)), and AM associates a variable (also representing an entity) with an adjective modifying it (as in (Jurafsky and Martin, 2000)). Binary predicates came from prepositions, except com(x145, x155). This is originated because in this particular syntactic context *há* means com (*with*), as stated in section 3.2.

4 Evaluation

4.1 Corpus Edite

The Edite corpus was collected after the Edite's project (da Silva, 1997; Reis et al., 1997), containing 680 questions about 68 tourist resources – from hotels to restaurants, golf fields, etc. It was

[30] We put some order in the variables generation: the position of the word in the text is the associated variable index.
[31] *Where are located the best hotels in Algarve*.
[32] *Where can one swim in the Green Coast?*.
[33] *How much*.
[34] Respectively, *What time*, *In which days*, *In which year*.
[35] *which are the beaches in Alentejo, with a blue flag and having life-guard?*.

built by a group of ten people, being each one responsible for 68 questions concerning each tourist resource.

Edite is not a corpus naturally built, *i.e.*, it was not created by tourists in a real situation of information request. Nevertheless, it was a starting point for the presented translating system. In this way, whenever was important to distinguish semantic behavior about a given element or group of elements, we used this corpus. In fact, this corpus was the basis for the observations that allowed us to determine if there were different syntactic contexts associated to those different semantic behaviors.

4.2 *Onde* – disambiguation results

The Edite corpus contains 122 occurrences of the word *onde*, distributed as follows:

- 72 occurrences at the beginning of the question – that is, applying H1 allows to obtain 72 onde_int;
- 2 occurrences at the end of the question – as so, applying H2 allows to obtain 2 additional onde_int;
- 3 occurrences of *onde* (that do not occur at the beginning or at the end of the question) are followed by *é que* – meaning that we have more 3 onde_int;

Moreover,

- H4 and H5 allows to detect 37 onde_rel;
- H6 detects 7 additional onde_int.

Therefore, we were able to detect (correctly) the category of the word *onde* in 121 of the 122 cases. The remaining case is

Igreja Nossa Senhora do Rosário, onde fica?[36]

Enriching H2 – that is, adding that *onde* followed by *fica(m), encontra(m)* or *é (são)* in the last position of the statement is a onde_int – allows to disambiguate the remaining case.

It should be noticed that the order of application of these heuristics is relevant. Consider, for example, the sentence

Onde é que há lagos onde possa fazer ski aquático?[37]

The first *onde* is disambiguated by H1. Only the fact that category *onde_int* is given to this *onde*, allows H3 to be applied and to disambiguate the second *onde*.

4.3 Identifying what is asked

100 questions were extracted from Edite's corpus and applied to JaVaLI! chain. Then, these sentences were manually treated, assuming correct arrowing relations. ASdeCopas was then applied and able to correctly identify 95% of what was asked. That is, in 95% of the cases, ASdeCopas was able to identify if we were asking for a location of an hotel, for the existence of a lake, for the possibility of visiting a museum, and so on. The remaining cases result from the incomplete treatment of the interrogative adverb *como* (ex: *Como que se pode ir para o Castelo de Palmela?*[38]).

A final note goes to the possibility of evaluating ASdeCopas' performance incrementally: as rules are independent from each other, one can evaluate the results produced by a given set of rules – for example, rules regarding interrogative particles – without checking if the system already has rules for verbs, adverbs, etc.

5 Conclusions

We presented JaVaLI!, a system that tries to translate unrestricted text into a formal language. Special focus was given to linguistic clues that can help to disambiguate words and also to make semantically equivalent sequences converge into the same formula. We showed some preliminary results of applying JaVaLI! to a corpus of 680 Portuguese questions.

Hypothetically, JaVaLI! could be applied to the semantic web as a translator, trying to transform existing HTML sources, for example, into XML/RDF. However, until now, we made no effort in that direction. In fact, more promising is to use JaVaLI! as a translator of NL queries into

[36] *Nossa Senhora do Rosrio church, where is it?*

[37] *Where are the lakes where one can ski?*

[38] *How can one go to Palmela's Castle?*

a semantic web query language as for example TRIPLE (Sintek and Decker, 2001).

References

Steven Abney. 1995. *Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax*. CSLI.

Salah Aït-Mokhtar. 1998. *L'analyse Présyntaxique en une seule étape*. Ph.D. thesis, Université Blaise Pascal, Feb.

James Allen. 1995. *Natural Language Understanding (second edition)*. The Benjamin Cummings Publishing Company, Inc.

Fernando Batista and Nuno Mamede. 2002. SuSAna: Módulo multifuncional da análise sintáctica de superfície. In Julio Gonzalo, Anselmo Peñas, and Antonio Ferrández, editors, *Proc. Multilingual Information Access and Natural Language Processing Workshop (IBERAMIA 2002)*, pages 29–37, Sevilla, Spain, November.

Gabriel Bès and Caroline Hagège. November, 2001. Properties in 5P (soon in the GRIL pages).

Gabriel G. Bès. 1999. La phrase verbal noyau en français. In *in Recherches sur le français parlé, 15*, pages 273–358. Université de Provence, France.

Luísa Coheur and Nuno Mamede. 2003. ASdeCopas: a syntactic-semantic interface. paper submited to EPIA 2003.

Luísa Coheur. 2003. AlgAS, um algoritmo de pré-análise semântica. Technical Report RT/008/02-CDIL, L^2F-Laboratório de Sistemas de Língua Falada, Inesc-id, Lisboa, Portugal, Maro.

Luísa Marques da Silva. 1997. Edite, um sistema de acesso a base de dados em linguagem natural, análise morfológica, sintáctica e semântica (master thesis). Master's thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal.

Abbaci Faiza. 1999. Développement du module Post-SMorph. Master's thesis, Mémoire de DEA de linguistique et informatique, GRIL, Université Blaise Pascal, Clermont-Ferrand.

Caroline Hagège. 2000. *Analyse Syntatic Automatique du Portugais*. Ph.D. thesis, Université Blaise Pascal, Clermont-Ferrand, France.

Daniel Jurafsky and James Martin, 2000. *Speech and Language Processing*, chapter 15. Prentice Hall.

Boris Katz and Jimmy Lin. 2000. Rextor: A system for generating relations from natural language. In *Proceedings of the ACL 2000 Workshop of Natural Language Processing and Information Retrieval (NLP&IR)*, October.

Boris Katz, Deniz Yuret, Jimmy Lin, Sue Felshin, Rebecca Schulman, and Adnan Ilik. 1998. Blitz: A preprocessor for detecting context-independent linguistic structures. In *Proceedings of the 5th Pacific Rim Conference on Artificial Intelligence (PRICAI '98)*, November.

David Matos, Joana Paulo, and Nuno Mamede. 2003. Managing linguistic resources and tools. In *6th PROPOR Workshop*, Faro, Portugal, June.

David Milward. 1999. Towards a robust semantics for dialogue using flat structures. In *Proceedings of Amstelogue.*

Joana Lúcio Paulo. 2001. PAsMo - pós-análise morfológica. Relatório técnico, Instituto Superior Técnico, Lisboa.

M. Poesio. 1994. *Ambiguity, Underspecification and Discourse Interpretation*. "ITK, Tilburg University".

Paulo Reis, J. Matias, and Nuno Mamede. 1997. Edite - a natural language interface to databases: a new dimension for an old approach. In *Proceeding of the Fourth International Conference on Information and Communication Technology in Tourism (ENTER'97)*, Edinburgh, Escócia. Springer-Verlag.

Michael Sintek and Stefan Decker. 2001. Triple – a query language for the semantic web. In *ICEC2001, Workshop on Semantic Web-based E-Commerce and Rules Markup Languages*, November.