Puzzle Out the Semantic Web Search

Dora Melo¹, Irene Pimenta Rodrigues², and Vitor Beires Nogueira²

Instituto Politécnico de Coimbra, Portugal dmelo@iscac.pt,
Universidade de Évora, Portugal {ipr,vbn}@di.uevora.pt

Abstract. In this paper we propose a Cooperative Question Answering System that takes as input natural language queries and is able to return a cooperative answer based on semantic web resources, more specifically DBpedia represented in OWL/RDF as knowledge base and WordNet to build similar questions. Our system resorts to ontologies not only for reasoning but also to find answers and is independent of prior knowledge of the semantic resources by the user. The natural language question is translated into its semantic representation and then answered by consulting the semantics sources of information. The system is able to clarify the problems of ambiguity and helps finding the path to the correct answer. If there are multiple answers to the question posed (or to the similar questions for which DBpedia contains answers), they will be grouped according to their semantic meaning, providing a more cooperative answer to the user.

Keywords: Natural Language, Ontology, Question Answering, Semantic Web

1 Introduction

Ontologies and the semantic web [1] became a fundamental methodology to represent the conceptual domains of knowledge and to promote the capabilities of semantic question answering systems [2]. These systems by allowing search in the structured large databases and knowledge bases of the semantic web can be considered as an alternative or as a complement to the current web search.

There is a gap between users and the semantic web: it is difficult for endusers to understand the complexity of the logic-based semantic web. Therefore it is crucial to allow a common web user to profit from the expressive power of semantic web data models while hiding its potential complexity. There is a need for user-friendly interfaces that scale up to the web of data and support end-users in querying this heterogeneous information source.

Consistent with the role played by ontologies in structuring semantic information on the web, ontology-based question answering systems allows us to exploit the expressive power of ontologies and go beyond the usual "keyword-based queries".

Question answering systems provide concise answers to natural language question posed by users in their own terminology [3]. Those answers must also be in natural language in order to improve the system and provide a better user friendly interface.

In this paper we propose a cooperative question answering system that receives queries expressed in natural language and is able to return a cooperative answer, also in natural language, obtained from resources on the semantic web (Ontologies and OWL2 Descriptions). The system starts a dialogue whenever there is some question ambiguity or when it detects that the answer is not what the user expected. Our proposal includes deep parsing³, use of ontologies and other web resources such as the WordNet [4] and the DBpedia [5].

Our goal is to provide a system that is independent of prior knowledge of the semantic resources by the user and is able to answer cooperatively to questions posed in natural language. The system maintains the structure of the dialogue and this structure provides a context for the interpretation of the questions, includes implicit content such as spatial and temporal knowledge, entities and information useful for the semantic interpretation, like discourse entities used for anaphora resolution, on finding what an instance of an expression is referring to.

This paper is organized as follows. First, in Section 2, we present a brief overview on cooperative question answering. In Section 3, we introduce the proposed system, describing the main components of its architecture. In parallel, we present an example as an illustration of the system functionality. Afterwards, in Section 4, we present related work, highlighting the main differences to the proposed system. Finally, in Section 5, we present the conclusions and the future work.

2 An Overview on Cooperative Question Answering

Question answering may be seen as the task of automatically answering a question posed in natural language. To find the answer to a question, a question answering system may use either a pre-structured database or a collection of natural language documents. Therefore, a question answering system provides precise answers to user questions by consulting its knowledge base.

In order to provide users with accurate answers, question answering systems need to go beyond lexical-syntactic analysis to semantic analysis and processing of texts and knowledge resources. Moreover, question answering systems equipped with reasoning capabilities can derive more adequate answers by resorting to knowledge representation and reasoning systems like Description Logic and Ontologies.

³ Deep parsing is directly based on property grammars. It consists, for a given sentence, in building all the possible subsets of overlapped elements that can describe a syntactic category. A subset is positively characterized if it satisfies the constraints of a grammar.

A cooperative answer [6] to a query is an indirect answer that is more helpful to the user than a direct, literal answer would be. A cooperative answer may explain the failure of a query to produce results and/or suggest follow-up queries. In the case where a query does produce results, a cooperative answer may provide additional information not explicitly requested by the user. Cooperative answers arose in the context of natural language question answering and they were originally motivated by the desire to follow the conventions of human conversation in human machine interactions performed in natural language.

Advanced reasoning for question answering systems raises new challenges since answers are not only directly extracted from texts or structured databases but also constructed via several forms of reasoning in order to generate answer explanations and justifications. Integrating knowledge representation and reasoning mechanisms allow, for example, to respond to unanticipated questions and to resolve situations in which no answer is found in the data sources. Cooperative answering systems are typically designed to deal with such situations by providing useful and informative answers. These systems should identify and explain false presuppositions or various types of misunderstandings found in questions.

3 Proposed System

Very briefly, the proposed system receives a natural language question and translates into a semantic representation using Discourse Representation Structures⁴ (DRS). Then, after consulting the semantics sources of information, provides a natural language answer. If there are multiple answers to the question posed (or to the similar questions for which DBpedia contains answers), they will be grouped according to their semantic meaning, providing a more cooperative, informative and clear answer to the user. Therefore, we consider that our system provides a user friendly interface.

The framework to develop our system was Prolog with several extensions and libraries. Among the reasons for such choice is the fact that there is a wide range of libraries for querying and processing of ontologies OWL2, WordNet has an export for Prolog and there are extensions that allow us to incorporate the notion of context into the reasoning process. Moreover, Wielemaker [7] provides a study for query translation and optimization more specifically the SeRQL RDF query language, where queries are translated to Prolog goals, optimized by reordering literals. Finally, in [8] the authors describe how to develop a semantic web application entirely in Prolog.

Our system architecture is presented in Figure 1 and to help its understanding we describe the main components in the following subsections.

⁴ For us a DRS is a set of referents, universally quantified variables and a set of conditions (first-order predicates). The conditions are either atomic (of the type $P(u_1,...,u_n)$ or $u_1=u_2$) or complex (negation, implication, disjunction, conjunction or generalized quantifiers).

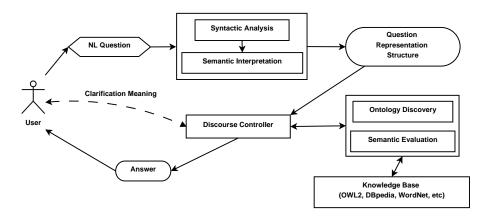


Fig. 1. Question Answering System Architecture.

3.1 Semantic Interpretation

Semantic analysis (or interpretation) is built using first-order logic [9] extended with generalized quantifiers [10]. We take special care with the discourse entities in order to have the appropriate quantifier introduced by the determinant interpretation. At this step, the syntactic structure of the question is rewritten into a DRS, that is supported by Discourse Representation Theory [11].

The implementation of this component follows an approach similar to the one for constructing a question answering system over documents databases proposed in [12]. The system consists of two separated modules: preliminary analysis of the documents (information extraction) and processing questions (information retrieval). The system is looking for processing the corpus and the questions, supported by theories of computational linguistics: syntactic analysis (grammatical restrictions) using deep parsing, followed by semantic analysis using the theory of discourse representation and finally the semantic (pragmatic) interpretation using ontology and logical inference.

As an illustration, consider the question "All French romantic writers have died?". The syntactic analysis generates a derivation tree, obtained from grammatical interpretation, that is rewritten according to a set of rules and integrated into a DRS, expressed in Prolog facts. In our study, it is stated by the following representation structure:

where the referent of the discourse is all-X, with X an universally quantified discourse entity, the main predication of the question is die(X) and the presupposed predications are writer(Y), french(Y), romantic(Y), with Y an existential quantified discourse entity. The system has to find all entities X that must verify the main predication condition only for those entities Y that verify

all the question presupposed conditions. The answer to the question will be a list with all french, romantic, writers resource entities who died.

3.2 Ontology Discovery

The Ontology Discovery is guided by the Discourse Controller to obtain the extension of sentence representation along with the reasoning process. The reasoning context and the question meaning will change whenever the Discourse Controller reaches a dead end.

This system module looks for similarities between labels according to their string-based, taking into account abbreviations, acronyms, domain and lexical knowledge. To maximize recall, the ontology search looks for classes, properties or instances that have labels matching a search term either exactly or partially and, if an answer is not achieved, each term in the query is extended with its synonyms, hypernyms and hyponyms obtained from WordNet [13]. Afterwards we extract a set of semantic resources which may contain the information requested.

Continuing the example of the previous section, in order to obtain the extension of sentence representation along the reasoning process, the system has to find the classes, properties or instances that have labels matching the search terms 'writer', 'french', 'romantic' and 'died', either exactly or partially. So, the system has to find the answers to the following questions:

- Which Classes, Properties or Instances represent the concept 'writer'? The system finds the DBpedia class Writer ⁵, with property domain Work and domain range Person. These domains inform the system of the class properties and can confirm whether this is related with the question, if not will be thrown away and a new search will be made. For instance, at the grammatical interpretation step, one of the presupposition found was that the entities that verify the question have to be persons. So, if the class Writer, does not have a relation with the class Person, or can't be applied to persons, at the phase of semantic interpretation it wouldn't be added to the set of facts that represent the information provided by the question and wouldn't be considered in the construction of the answer;
- Which Classes, Properties or Instances represent the concept 'french'? The DBpedia has a class birthPlace ⁶ (an entity of type ObjectProperty, with property domain Person and domain range Place) that represents the place where some person was born. The term 'french' is also interpreted as a "person of France" and have as a direct hypernym the term "country" (obtain from WordNet), so the system also has to find the classes, properties or instances of all similar meanings to the initial term that could lead the system to the correct answer;

⁵ http://dbpedia.org/ontology/Writer

⁶ http://dbpedia.org/ontology/birthPlace

- Which Classes, Properties or Instances represent the concept 'romantic'?
 The system finds the DBpedia resource Romanticism ⁷ (an entity of type Thing, an instance of property movement ⁸);
- Which Classes, Properties or Instances represent the concept 'die'? The DBpedia has a class deathDate ⁹ (an entity of type DatatypeProperty, with property domain Person and domain range date) that represents the death date of a person. The relation between the terms 'die' and 'death' can be made by searching the WordNet, where the term 'die' can be interpreted as a "decease", that in turn have as synonym the term "death".

The next step is the construction of query(ies) needed to verify the initial question. If the question doesn't have an answer, a set of similar questions is constructed. Querying the WordNet, the system obtains similar terms to those that compose the initial question. This set of similar questions will enrich the knowledge domain and helps the interpretation of the original question or in the construction of its answer. If this set of new questions leads the system to different answers, we are in the presence of an ambiguity and the user is invoked to clarify it. If the system did not find any correspondence to a word and its derivatives, the user is informed and can clarify the system by reformulating the question or presenting others query(ies).

3.3 Semantic Evaluation

Semantic evaluation is intended to be the pragmatic evaluation¹⁰ step of the system, where the question semantic is transformed into a constraint satisfaction problem. This is achieved by adding conditions that constrain the discourse entities. Moreover, this extra information (regarding the question interpretation) can help the Discourse Controller to formulate a more objective answer.

The semantic evaluation must reinterpret the semantic representation of the sentence based on the ontology considered in order to obtain the set of facts that represent the information provided by the question. Therefore, the process responsible for the semantic evaluation receives the DRS of the question and interprets it in a knowledge base with rules derived from the ontology and the information contained in the knowledge base like DBpedia and WordNet.

Back to our example, to solve the constraint problem the Dialogue Controller generates and poses questions such "Who are the French romantic writers?" to the question answering system, whose representation structure is

⁷ http://dbpedia.org/page/Romanticism

⁸ http://dbpedia.org/property/movement

⁹ http://dbpedia.org/ontology/deathDate

The pragmatic evaluation is the capability of to judge or calculate the quality, importance, amount or value of problems solutions that are solved in a realistic way which suits the present conditions rather than obeying fixed theories, ideas or rules.

First and according to the domain knowledge, the interpreter will transform the conditions of the DRS into OWL, i.e., constructs the relative condition based in the ontology. For instance, the condition ontology_writer will represent the DRS condition writer. Therefore, the new representation structure¹¹ for the question is

After obtaining this new set of DRS, the terms of the ontology will be interpreted as usual Prolog predicates. Then, by applying the unification mechanism of Prolog the system will obtain the answer to the question. So the answer to initial question will be

Francois-Rene de Chateaubriand (1768-1848) Alphonse de Lamartine (1790-1869) Alfred de Musset (1810-1857) Victor Hugo (1802-1885) Henri-Marie Beyle, Stendhal (1783-1842)

3.4 Discourse Controller

The Discourse Controller is a core component that is invoked after the natural language question has been transformed into its semantic representation. Essentially the Discourse Controller tries to make sense of the input query by looking at the structure of the ontology and the information available on the semantic web, as well as using string similarity matching and generic lexical resources (such as WordNet).

In Figure 2, we represent the architecture of the Discourse Controller. In outline, after transforming the natural language question into its semantic representation the Discourse Controller is invoked and controls all the steps until the end, i.e until the system can return an answer to the user. More specifically, the Ontology Discover is invoked in order to provide the extension of sentence representation. If the ontology representation of a term is not found, the Discourse Controller is alerted and the user is called to clarify it. When the extension of the sentence representation is complete, the Discourse Controller adds to his knowledge a set of semantic resources. Afterwards, the Semantic Evaluation is invoked. In this step, the question semantic is transformed into a constraint satisfaction problem, by adding conditions that constraint the discourse entities. This extra information can help the Discourse Controller to formulate a more

¹¹ The condition ontology_term represents the class, property or instance in the ontology that is the meaning of the term. If the interpreter has more than one possible ontology conditions for each term then will get several DRS rewritten with the terms of the ontology.

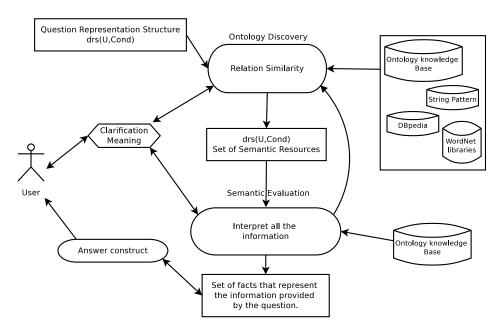


Fig. 2. Discourse Controller Module Architecture.

objective answer. If in the interpretation of all the information leads the Discourse Controller to an empty answer or to multiple answers, the user is called to clarify it and may be necessary to re-invoke the Ontology Discover. The process is finalized when the Discourse Controller is able to return an answer to the question posed by the user.

The Dialogue Controller deals with the set of discourse entities and is able to compute the question answer. It has to verify the question presupposition, choose the sources of knowledge to be used and decide when the answer has been achieved or to iterate using new sources of knowledge. The decision of when to relax a question in order to justify the answer and when to clarify a question and how to clarify it also taken by in this module.

Whenever the Discourse Controller isn't sure how to disambiguate between two or more possible terms or relations in order to interpret a query, it starts a dialogue with the user and asks him for disambiguation. The clarification done by the user will be essential for the Discourse Controller, this way obtaining the right answer to the query posed by the user. For instance, the question "Where is the Taj Mahal?", 'Taj Mahal' could be mapped into the name of a Mausoleum, a Casino Hotel or an Indian Restaurant and only the user can clarify about the intended meaning. The more cooperative and interactive the Discourse Controller is, the closer it will be to the correct answer.

Another important aspect of the Discourse Controller is to provide a friendly answer to the user. The answer should be as close as possible to the natural language. For instance, the question answering system has to respond "yes"

or "no" when the user posed the query "Is Barack Obama the President of the USA?". In this case, the answer will be "yes". However, the answer must be more informative for the user. Some concepts are defined in the temporal context, even if implicitly, and the answer should be more clear and informative. For instance, the term 'President', in the context of the question, is defined as the title of head of state in some republics and has an associated duration for the mandate, a start date (date of election, date on taking office), and an end date of the mandate. So the answer to the question "Is Barack Obama the President of the USA?" should be "Yes, Barack Obama is the actual President of USA", that is more cooperative and informative.

For the cases where the answer to a question of type Yes/No is "No", the Discourse Controller will return a complete answer, clarifying the negation. If we consider the question "All the capitals of Europe have more than 200,000 inhabitants?" that has a "No" as an answer, the system will construct the proper answer that clarify the user and will return "No, 9 capitals of Europe have less than or equal to 200,000 inhabitants".

If there are multiple answers to the question posed by the user (or to the similar questions for which DBpedia contains answers), they will be grouped according to their semantic meaning, providing a more cooperative and clean answer to the user. To do so, the discourse controller has to reason over the question and construct the answer, well constructed questions have always the right words that help in the answer construction. For the question "Where is the Taj Mahal?", the user is called to clarify the system about the ambiguity of the question: Taj Mahal is a Mausoleum, a restaurant or Casino Hotel; and consider that the user is not able to clarify it or he simply wants that the system returns all possible answers. So when the system has all the answers to all possible interpretations for the question posed by the user, the Discourse Controller will not list the answer in a random way, but will list the answer according to their semantic meaning:

```
Mausoleum Taj Mahal is in Agra, India
Casino hotel Taj Mahal is in Atlantic City, NJ, USA
Indian Restaurant Taj Mahal is in New Farm, Brisbane, Australia
Indian Restaurant Taj Mahal is in 7315 3rd Ave. - Brooklyn, NY, USA
```

Our dialogue system has as main objective the use of interaction to obtain more objective and concrete answers. It is not used only to clarify the problems of ambiguity, but also to help finding the path to the correct answer. Making the dialogue system more cooperative makes one able to get closer to the answer desired by the user. In many cases, the user is the only one who can help the system in the deduction and interpretation of information.

4 Related Work

The representation of questions with generalized quantifiers as in [14] allows the use of various natural language quantifiers like all, at least 3, none, etc. Moreover, the question evaluation also resorts to logic programming with constraints.

A query language for OWL based on Prolog is presented in [15]. The author proposes a way of defining a query language based on a fragment of Description Logic and a way of mapping it into Prolog by means of logic rules.

An illustration of a question answering system for the Portuguese language that uses the web as a database, through meta-search on conventional search engines can be seen in [16]. This system uses surface text patterns to find answers in the documents returned by search engines. Another example of a question answering system where domain knowledge is represented by an ontology can be found in [17]: it is presented an interface system for question answering Chinese natural language that runs through a natural language parser. In [18], the author illustrate the capabilities for cooperative response generation implemented in Artificial Intelligence systems.

In [19] we find a declarative approach to represent and reason about temporal contextual information. In this proposal each question takes place in a temporal context and that context is used to restrict the answer.

The fundamental techniques for computing semantic representations for fragments of natural language and performing inference with the result are presented in [20]. The primary tools used are first-order logic and lambda calculus, where all the techniques introduced are implemented in Prolog. The authors also show how to use theorem provers and model builders in parallel to deal with natural language inference.

PowerAqua [21] is a multi-ontology-based question answering system that takes as input queries expressed in natural language and is able to return answers drawn from relevant distributed resources on the semantic web. PowerAqua allows the user to choose an ontology and then ask natural language queries related to the domain covered by the ontology. The system architecture and the reasoning methods are completely domain-independent, relying on the semantics of the ontology, and the use of generic lexical resources, such as WordNet.

An overview of cooperative answering in databases is presented in [22]. A logic-based model for an accurate generation of intensional responses within a cooperative question answering framework is proposed by the author of [23]. The author developed several categories of intensional forms and a variabledepth intensional calculus that allows for the generation of intensional responses at the best level of abstraction and shows that it is possible to generate natural responses on a template basis. The same author in [24], presents an approach for designing a logic based question answering system, WEBCOOP, that integrates knowledge representation and advanced reasoning procedures to generate cooperative responses to natural language queries on the web. This project was developed on a relatively limited domain that includes a number of aspects of tourism (transportation) and requires the development of a knowledge extractor from web pages (similarly to a knowledge extractor operating on passages resulting from an information retrieval component) and the elaboration of a robust and accurate question parser. The responses provided to users are built in web style by integrating natural language generation techniques with hypertexts in order to produce dynamic responses. Natural language responses are produced from semantic forms constructed from reasoning processes.

Our proposal is a friendly, simple and cooperative question answering system. The main difference is the cooperative way that it answers the natural language questions posed by the user. We interact with the user in order to disambiguate and/or to guide the path to obtain the correct answer to the query posted, whenever this is possible to do by the reasoner. We also use cooperation to provide more informed answers. The answers is presented in natural language and have to clarify what the system can infer about the question from the knowledge domain. Therefore, the cooperative answer provided by our system has to explain the failure of a query to produce results and/or suggest follow-up queries. In the case where a query does produce results, the cooperative answer will provide additional information not explicitly requested by the user.

5 Conclusions and Future Work

We presented a cooperative semantic web question answering system that receives queries expressed in natural language and is able to return a cooperative answer, also in natural language, obtained from semantic web resources (Ontologies and OWL2 Descriptions). The system is able of dialoguing when the question has some ambiguity or when it detects that the answer is not what user expected. Our proposal includes deep parsing and the use of ontologies and other web resources such as the WordNet and the DBpedia.

As future work, we intend to answer questions that are more elaborate and/or more difficult. Moreover, we also plan to extend to the Portuguese natural language. For this purpose it will be necessary to enrich the knowledge domain with concepts that may be deduced from the initial domain. Although the system is intended to be domain independent, it will be tested in a number of domains, with special relevance to the wine and the movies, since for these fields there are many resources available in the semantic web. We also plan to build a DRS generator, that builds the question semantics and retains additional information that allows the Discourse Controller to provide a more adequate cooperative answer. We contemplate about enlarging the knowledge base with other ontologies in order to support open domain question answering and take advantage of the vast amount of heterogeneous semantic data provided by the semantic web.

References

- 1. Horrocks, I.: Ontologies and the semantic web. Communications of the ACM ${\bf 51}$ (2008) ~58
- Guo, Q., Zhang, M.: Question answering based on pervasive agent ontology and Semantic Web. Knowledge-Based Systems 22 (2009) 443–448
- 3. Hirschman, L., Gaizauskas, R.: Natural language question answering: The view from here. Natural Language Engineering 7 (2001) 275–300
- 4. Fellbaum, C.: WordNet: An electronic lexical database. The MIT press (1998)

- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J.: Dbpedia: A nucleus for a web of open data. The Semantic Web 4825 (2007) 722–735
- Corella, F., Lewison, K.: A brief overview of cooperative answering. Journal of Intelligent Information Systems 1 (2009) 123–157
- Wielemaker, J.: An optimised Semantic Web query language implementation in Prolog. Logic Programming (2005) 128–142
- Wielemaker, J., Hildebrand, M., van Ossenbruggen, J.: Using Prolog as the fundament for applications on the semantic web. Proceedings of ALPSWS2007 (2007) 84–98
- 9. Hodges, W.: Classical logic I: first-order logic. The Blackwell guide to philosophical logic (2001) 9–32
- Barwise, J., Cooper, R.: Generalized quantifiers and natural language. Linguistics and philosophy 4 (1981) 159–219
- 11. Kamp, H., Reyle, U.: From Discourse to Logic. Volume 42 of Studies in Linguistics and Philosophy. Kluwer (1993)
- Quaresma, P., Rodrigues, I., Prolo, C., Vieira, R.: Um sistema de Pergunta-Resposta para uma base de Documentos. Letras de Hoje 41 (2006) 43–63
- 13. Witzig, S., Center, A.: Accessing wordnet from prolog. Artificial Intelligence Centre, University of Georgia (2003) 1–18
- 14. Rodrigues, I., Quintano, L., Ferreira, L.: Nl database dialogue question-answering as a constraint satisfaction problem. In: 18th Intl. Conf. on Applications of Declarative Programming and Knowledge Management (INAP'09), Univ. Évora (2009)
- Almendros-Jiménez, J.M.: A Prolog-based Query Language for OWL. Electronic Notes in Theoretical Computer Science 271 (2011) 3–22
- Rabelo, J., Barros, F.: Pergunte! uma interface em português para perguntaresposta na web. Master's thesis, Informatics Center, Federal University of Pernambuco, Brazil (2004) 1114–1117
- 17. Guo, Q.: Question answering system based on Ontology. In: Intelligent Control and Automation, 2008. WCICA 2008. 7th World Congress on, IEEE (2008) 3347–3352
- 18. Wahlster, W.: Cooperative access systems. Future Generation Computer Systems ${f 1}$ (1984) 103–111
- 19. Nogueira, V., Abreu, S.: Temporal contextual logic programming. Electronic Notes in Theoretical Computer Science 177 (2007) 219–233
- Blackburn, P., Bos, J.: Representation and inference for natural language: A first course in computational semantics. Center for the Study of Language and Information (2005)
- 21. Lopez, V., Motta, E.: Poweraqua: Fishing the semantic web. Semantic Web: Research and Applications (2006)
- Minker, J.: An overview of cooperative answering in databases. Flexible Query Answering Systems (1998) 282–285
- Benamara, F.: Generating intensional answers in intelligent question answering systems. Natural Language Generation (2004) 11–20
- 24. Benamara, F.: Cooperative question answering in restricted domains: the WE-BCOOP experiment. In: Proceedings of the Workshop Question Answering in Restricted Domains, within ACL. (2004)