Ecological Informatics 30 (2015) 328–344

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

An authoring tool for decision support systems in context questions of ecological knowledge

Antonio Ferrández [a,1], Jesús Peral [a,1], Elisa De Gregorio [a,1], Juan Trujillo [a,1], Alejandro Maté [a,1], Luis José Ferrández [a,1], Yenory Rojas [b,1]

[a] Dept. of Software and Computing Systems, University of Alicante, Carretera San Vicente S/N, Alicante, 03080, Spain
[b] Universidad Hispanoamericana, Llorente de Tibás, San José, Costa Rica

ARTICLE INFO

Article history:
Received 9 January 2015
Received in revised form 24 June 2015
Accepted 1 September 2015
Available online 10 September 2015

Keywords:
Decision support system
Natural language interface
Context questions
Ellipsis
Anaphora
Ontologies

ABSTRACT

Decision support systems (DSS) support business or organizational decision-making activities, which require the access to information that is internally stored in databases or data warehouses, and externally in the Web accessed by Information Retrieval (IR) or Question Answering (QA) systems. Graphical interfaces to query these sources of information ease to constrain dynamically query formulation based on user selections, but they present a lack of flexibility in query formulation, since the expressivity power is reduced to the user interface design. Natural language interfaces (NLI) are expected as the optimal solution. However, especially for non-expert users, a real natural communication is the most difficult to realize effectively.

In this paper, we propose an NLI that improves the interaction between the user and the DSS by means of referencing previous questions or their answers (i.e. anaphora such as the pronoun reference in "What traits are affected by them?"), or by eliding parts of the question (i.e. ellipsis such as "And to glume colour?" after the question "Tell me the QTLs related to awn colour in wheat"). Moreover, in order to overcome one of the main problems of NLIs about the difficulty to adapt an NLI to a new domain, our proposal is based on ontologies that are obtained semi-automatically from a framework that allows the integration of internal and external, structured and unstructured information. Therefore, our proposal can interface with databases, data warehouses, QA and IR systems. Because of the high NL ambiguity of the resolution process, our proposal is presented as an authoring tool that helps the user to query efficiently in natural language. Finally, our proposal is tested on a DSS case scenario about Biotechnology and Agriculture, whose knowledge base is the CEREALAB database as internal structured data, and the Web (e.g. PubMed) as external unstructured information.

1. Introduction and motivation

Decision support systems (DSS) usually require the access to huge resources of information. As the amount of information available globally on the Web and locally in intranets or databases keeps steadily growing, the necessity of mechanisms for effectively querying this information gains importance at the same pace (Cimiano and Minock, 2009). Moreover, with the wide availability of smart phones and tablets, the importance of intuitive ways of interacting with electronic devices has grown even more. Natural language interfaces (NLI) are an interesting option to interact with mobile devices due to their limited input and output functionality, which makes graphical interfaces less appealing (Popescu et al., 2003). Clearly, automatic speech recognition is a crucial component towards leveraging the use of NLIs.

Nowadays, the large amount of information obtained by scientific research in Life Sciences is stored in specialized databases (DBs), particularly about Genetics and Biotechnology (Matos et al., 2010). These huge DBs require optimized search strategies in order to extract biological information in a comfortable and efficient way by the user (Altman et al., 2008; Jensen et al., 2006). In this regard, there is a need to design simple interfaces which work with complex Molecular Biology concepts and ease the biologists the comprehension of the data. As Li et al. (2007) concludes, database query languages (e.g. SQL) can be intimidating to the non-expert, leading to the immense recent popularity for keyword based search or graphical interfaces in spite of their significant limitations. Cimiano and Minock (2009) describe different paradigms proposed in the past for querying information collections, among them form filling, query-by-example or menu-based approaches, as well as NLIs, relying either on controlled language or on more or less free language input. Obviously, Natural Language (NL) searching probably constitutes the most flexible and effective approach for interrogating a biological DB (Jamil, 2012). Since many biological DBs (e.g. GenBank or UCSC Genome Browser) employ graphical interfaces that are not prepared for arbitrary

E-mail addresses: antonio@dlsi.ua.es (A. Ferrández), jperal@dlsi.ua.es (J. Peral), edg12@alu.ua.es (E. De Gregorio), jtrujillo@dlsi.ua.es (J. Trujillo), amate@dlsi.ua.es (A. Maté), ljfp1@alu.ua.es (L.J. Ferrández), yrojas@uh.ac.cr (Y. Rojas).
[1] Tel.: 34 96 590 3400.

http://dx.doi.org/10.1016/j.ecoinf.2015.09.007
1574-9541/

questioning, previous effort has been made to build NL strategies oriented to Biomedicine and Biotechnology DBs (e.g. Clegg and Shepherd, 2007; Distelhorst et al., 2003; Goldsmith et al., 2009; Jamil, 2012). As Cimiano and Minock (2009) state, while the querying paradigm based on NL is generally deemed to be the most intuitive from a usage point of view, it has also been shown to be the most difficult to realize effectively. The main reasons for this difficulty are that:

- NL understanding is indeed a very difficult task due to ambiguities arising at all levels of analysis: morphological, lexical, syntactic, semantic, and pragmatic.
- A reasonably large grammar is required for the system to have an acceptable coverage.
- The NLI needs to be accurate.
- The system should be adaptable to various domains without a significant effort.

Therefore, there is still much work to be done in the field of NLIs. Our proposal and the case study in which it is proved lie on the query of biological knowledge by means of an NLI that handles the context of previous questions and resolves the ellipsis and anaphora ambiguity. For example, in the questions below, the second one needs to resolve the ellipsis to determine the aim of the question: "What QTLs are related to frost tolerance in durum wheat?" The third one needs to resolve the anaphor "these" from the context of the previous questions and their answers.

1 What QTLs are related to frost tolerance in barley?
2 In durum wheat?
3 What other traits are related to these?

The field of Genetic Engineering is an emergent discipline, which has expanded to biomedicine, agriculture and other related domains (Aleksejeva, 2014). Our case study focuses on a Plant Biotechnology industry, whose main target is to create Genetically Modified Organisms (GMOs). According to the World Health Organization, GMOs are "organisms in which the genetic material (DNA) has been altered in such a way that does not occur naturally" (World Health Organization, 2002) and they are obtained by inserting sequences of DNA from one organism to another. This plant breeding strategy has an important role in the world market, reaching the point that, in 2014, five seed companies control 35% of the global market (Le Buanec, 2008) and 33% of their product are GMOs (ETC, 2008; ISAAA, 2012; Meijerink and Danse, 2009; Rótolo et al., 2015). In cereals, fruits, vegetables, grains and legumes the production of modified seeds keeps growing.

In Plant Biotechnology, the QTL (Quantitative Trait Locus) analysis is highly useful, since allows us to identify the action, interaction, number, and precise location of the chromosomal regions containing one or more genes involved in specific phenotypic features (Falconer and Mackay, 1996; Kearsey, 1998; Lynch and Walsh, 1998; Miles and Wayne, 2008). Thereby, a QTL could be transferred in the laboratory from one organism to another for modifying one or more particular traits.

For this reason, to be updated with QTLs is especially important to design Genetic Engineering protocols, which improve plant varieties (i.e. enhancing its flavor and nutritional value, improving its cold resistance or producing fruits out of season). Given that our knowledge of the function of gene products is increasing rapidly, QTL databases try to collect all the relations between chromosome positions and biological features of many organisms.

Our proposal facilitates the decision-making process because it extends our previous work in Peral et al. (2015), in which, internal structured information (e.g. the CEREALAB database, Milc et al., 2011) and external unstructured data obtained from the Web (e.g. the PubMed URL) are integrated and presented to the user in a dashboard. Here, we extend the NLI in this previous work by handling the context of previous questions and their answers by resolving ellipsis and anaphora. For instance, the user could find in series of questions if a QTL is related to several phenotypical traits (pleiotropy), or if a trait is influenced by several QTLs (polygenic traits with multifactorial inheritance). Moreover, interesting commercial information could be retrieved, like the existence of transgenic varieties and their market price, in order to establish competitive prices for new transgenic seeds.

The paper is structured as follows. In Section 2, we summarize the most relevant related work. In Section 3, we introduce our proposal for the integrated "anaphora + ellipsis + context question" authoring tool for facilitating the decision-making process. In Section 4, in order to clarify our proposal, we illustrate the application of our proposal on the case study in which the CEREALAB database is queried. We conclude the paper with the summary of our main contributions and our directions for future works.

2. Related work

As we introduced in our previous motivation section, decision support systems integrate a variety of interfaces for querying databases or data warehouses. So far, these interfaces have been implemented as graphical systems, which ease to dynamically constrain query formulation based on user selections, in order to only build valid questions. However, these graphical interfaces present the disadvantages of a lack of flexibility in query formulation, since the expressivity power is reduced to the user interface design. Therefore, they provide less expressivity power than textual NL interfaces, as well as they force the user to learn an additional formal language or graphical system. Then, we could consider that NL interfaces are the optimal solution, but they also present some disadvantages such as the difficulty of dealing with NL, that is to say the linguistic coverage, ambiguity and the managing of discourse that allows a real natural way of communication (e.g. context questions, anaphora and ellipsis resolution).

The following four subsections discuss the state-of-the-art of these issues: NL interfaces, context questions, anaphora and ellipsis resolution.

2.1. Related work about NL interfaces

In this subsection, we review only those NL interfaces related to our proposal. That is why we group these interfaces according to the following: they deal with the problem of anaphora, ellipsis, and context questions, as summarized in Table 1. We should emphasize that most of these interfaces do not handle with any of these problems, such as: Popescu et al. (2003), Stratica et al. (2005), or Barbosa et al. (2006).

With regard to those interfaces that face anaphora problem, the work by Li et al. (2005) presents NaLIX, a generic interactive NL query interface to an XML database, which deals with query pronouns, showing a warning indicating the possible loss of search quality if incorrect anaphora resolution would be made. However, other kinds of anaphors or context questions are not resolved. In Laukaitis and Vasilecas (2007), the authors present an agent based NL dialog architecture for data querying from database management systems. In their architecture, the NL processing module implements morphology, syntax and lexical semantics analysis. The final result of those steps is identified as triplets: entities, relationships and associated probabilities. For this purpose, GATE system (Cunningham et al., 2000) has been used. However, ellipsis or context questions are not handled.

Concerning those interfaces that use ontologies as our proposal does, the work by Cimiano et al. (2008) presents the NLI named ORAKEL. It is an ontology-based NL system: (a) the ontology for a certain knowledge base is used to guide the lexicon construction process; (b) ORAKEL is an NLI which relies on deduction to answer a user's query. The main disadvantage of the system is that the domain lexicon needs to be handcrafted by a domain expert instead of an automatic process that