Querying Biomedical Ontologies in Natural Language using Answer Set Programming

Halit Erdogan[1], Umut Oztok[1], Yelda Erdem[2], and Esra Erdem[1]

[1] Faculty of Engineering and Natural Sciences, Sabancı University, İstanbul, Turkey
[2] Research and Development Department, Sanovel Pharmaceutical Inc., İstanbul, Turkey

Recent advances in health and life sciences have led to generation of a large amount of data. To facilitate access to its desired parts, such a big mass of data has been represented in structured forms, like biomedical ontologies. On the other hand, representing ontologies in a formal language, constructing them independently from each other and storing them at different locations have brought about many challenges for answering queries about the knowledge represented in these ontologies. One of the challenges for the users is to be able represent a complex query in a natural language, and get its answers in an understandable form: Currently, such queries are answered by software systems in a formal language, however, the majority of the users lack the necessary knowledge of a formal query language to represent a query; moreover, none of these systems can provide informative explanations about the answers. Another challenge is to be able to answer complex queries that require appropriate integration of relevant knowledge stored in different places and in various forms.

In this work, we address the first challenge by developing an intelligent user interface that allows users to enter biomedical queries in a natural language, and that presents the answers (possibly with explanations if requested) in a natural language. We address the second challenge by developing a rule layer over biomedical ontologies and databases, and use automated reasoners to answer queries considering relevant parts of the rule layer. The main contributions of our work can be summarized as follows:

- We introduce a controlled natural language, a subset of natural language with a restricted grammar and vocabulary, specifically for biomedical queries towards drug discovery; we call this controlled natural language as BIOQUERYCNL [4]. For instance, in this language, we can pose the following query:

  > "What are the genes that are targeted by the drug Epinephrine and that interact with the gene DLG4?"

- We present an algorithm that converts a biomedical query in BIOQUERYCNL into a program in answer set programming (ASP) — a formal framework to automate reasoning about knowledge [8] — making use of the parsing engine APE [5]. Figure 1 shows the overall idea behind this algorithm. For instance, according to this algorithm, the query above is translated into the following ASP program:

  ```
  what_be_genes(GN1) :-
     gene_gene(GN1,"DLG4"),
     drug_gene("Epinephrine",GN1).
  ```

  where `gene_gene` and `drug_gene` are defined in a "rule layer".

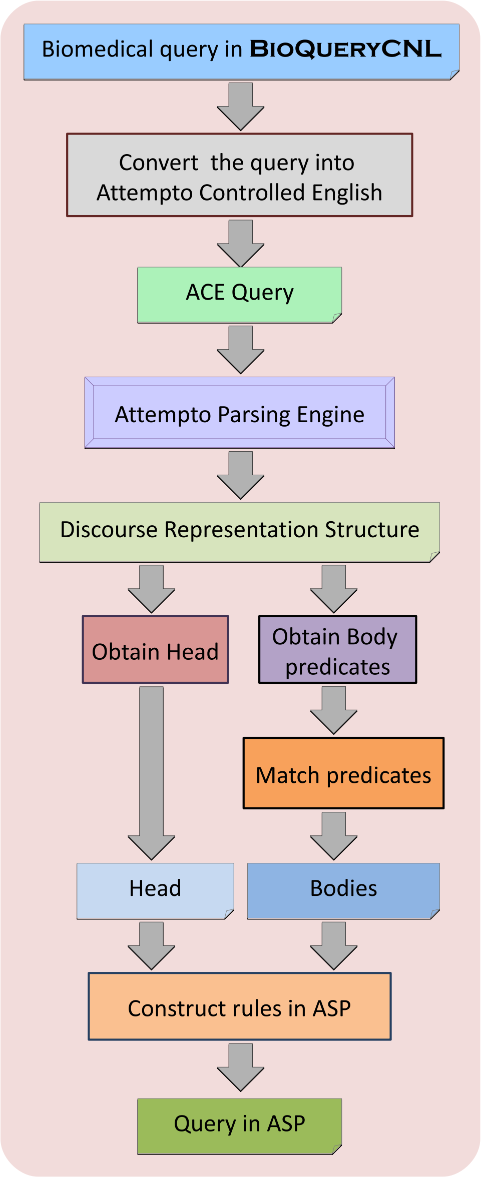

Fig. 1. Transforming a query in BioQueryCNL into an ASP program.

- Once we transform the biomedical query into an ASP program and extract the relevant part of the rule layer (also an ASP program), we can compute its answers (if exists) using a state-of-the-art ASP system, such as CLASP [6], DLV [3,7] or DLVHEX [2], as described in [1]. Figure 2 shows the overall idea behind this algo-

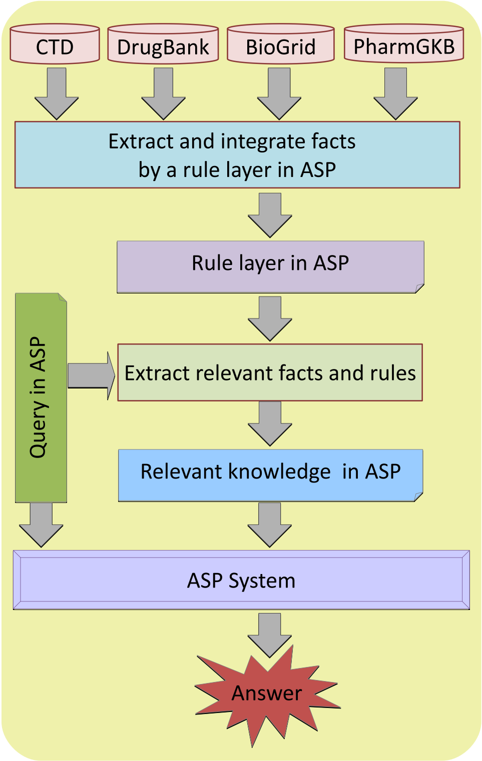
CTD
DrugBank
BioGrid
PharmGKB
Extract and integrate facts by a rule layer in ASP
Rule layer in ASP
Query in ASP
Extract relevant facts and rules
Relevant knowledge in ASP
ASP System
Answer

Fig. 2. Extracting and integrating knowledge from ontologies or databases that is relevant to a given query, and finding an answer to the query using an ASP system.

rithm. For instance, using CLASP, we compute the following answer to the query above: "ADRB1".

- We construct an algorithm to provide minimal explanations to the answers. For instance, for "ADRB1" our algorithm provides the following minimal explanation:

> the drug "Epinephrine" targets the gene "ADRB1" according to CTD and the gene "ADRB1" interacts with the gene "DLG4" according to BIOGRID.

The applicability of our methods is illustrated with some complex queries over PHARMGKB, DRUGBANK, BIOGRID, SIDER and CTD, using the ASP systems CLASP (with GRINGO), DLV and DLVHEX.

Acknowledgments. This work has been supported by TUBITAK Grant 108E229.

References

1. Bodenreider, O., Coban, Z.H., Doganay, M.C., Erdem, E.: A preliminary report on answering complex queries related to drug discovery using answer set programming. In: Proc. of the 3rd

International Workshop on Applications of Logic Programming to the Semantic Web and Web Services (2008)

2. Eiter, T., G.Ianni, R.Schindlauer, H.Tompits: Effective integration of declarative rules with external evaluations for Semantic-Web reasoning. In: Proc. of ESWC (2006)
3. Eiter, T., Leone, N., Mateis, C., Pfeifer, G., Scarcello, F.: A deductive system for non-monotonic reasoning. In: Proc. of LPNMR. pp. 364–375 (1997)
4. Erdem, E., Yeniterzi, R.: Transforming controlled natural language biomedical queries into answer set programs. In: Proc. of the Workshop on BioNLP. pp. 117–124 (2009)
5. Fuchs, N.E.: Attempto controlled english. In: Proc. of WLP. pp. 211–218 (2000)
6. Gebser, M., Kaufmann, B., Neumann, A., Schaub, T.: T.: Conflict-driven answer set solving. In: Proc. of IJCAI. pp. 386–392 (2007)
7. Leone, N., Pfeifer, G., Faber, W., Eiter, T., Gottlob, G., Perri, S., Scarcello, F.: The dlv system for knowledge representation and reasoning. ACM Trans. Comput. Log. 7(3), 499–562 (2006)
8. Lifschitz, V.: What is answer set programming? In: Proc. of AAAI (2008)