Hybrid Question Answering over Heterogeneous Data
(Internship Subject for Master 2 Research)

- Laboratories:

  LIMSI, CNRS, Université Paris-Saclay, France

  LRI, CNRS UMR 8623, Université Paris-Saclay, France

- Duration: April to Septembre, 2016 with possibility of continuing with a PhD position
- Start Date: 01/04/2016
- Supervisors: Brigitte Grau (LIMSI); Yue Ma (LRI); Pierre Zweigenbaum (LIMSI)

Description

More and more information on individuals (e.g., persons, events, biological objects) are available electronically in a structured or semi-structured form. However, selecting individuals satisfying certain constraints based on such data manually is a complex, error-prone, and time and personnel-consuming effort. For this reason, tools that can automatically or semiautomatically answer questions based on the available data need to be developed.

While simple questions can directly be expressed and answered using keywords in natural language, complex questions that can refer to type and relational information will increase the precision of retrieved results, and thus reduce the effort for posterior manual verification of the results. One example for this situation is the setting where electronic patient records are used to find patients satisfying non-trivial combinations of certain properties, such as eligibility criteria for clinical trials. For instance, assume that we are interested in finding male patients who are at risk for a heart attack. Instead of keywords, this can be better described by the structured conjunctive query:

$$\exists y.risk(x, y) \wedge Myocardial\_infarction(y) \wedge Male(x).$$

Now assume that the patient database stores the facts about Bob, represented below, that he has high blood pressure, is male, and has a history of hypertension:

$$\begin{array}{c} systolic\_pressure(BOB, P1),\ High\_pressure(P1), \\ history(BOB, H1),\quad Hypertension(H1),\ Male(BOB) \end{array}$$

Note that we cannot conclude that Bob is a desired patient directly from the above pieces of information. This is due to the lack of background knowledge of the medicine domain. Now suppose that we have an ontology (represented in Description Logics [1]) saying that patients with high blood pressure have hypertension and that patients who currently have hypertension and also have a history of hypertension are at risk for a heart attack:

$$\begin{array}{rcl} \exists systolic\_pressure.High\_pressure & \sqsubseteq & \exists finding.Hypertension \\ \exists finding.Hypertension \sqcap \exists history.Hypertension & \sqsubseteq & \exists risk.Myocardial\_infarction \end{array}$$

Given the facts about Bob and the ontology, now we can derive that Bob is an answer to the query. Obviously, the integration of ontologies is meaningful for complex query answering.

The context of this internship is therefore a nouvel answering question paradigm that integrates both formal database-like query answering and texts based answering by information extraction methods [2, 3, 6, 4, 5, 7]. While formal queries are powerful in representing complex questions and exploring background knowledge, they are often difficult to master, and such an advanced answering system cannot be used without a user adapted interface. Thus, our aim is to allow a user to formulate her need with natural

language questions that can be complex pieces of texts. Apart from an easy interface, natural language will enable to formulate constraints that cannot be represented formally due to the expressiveness limit of the formal language, but that can be directly verified using textual data. To this purpose, we need to combine answers to a formal query generated from a NL question with answers found based on information retrieval techniques, which is among the identified challenges in question answering systems [8]. The tasks of this internship consist in the following aspects:

- The first task will be a detailed review of the state-of-art on question answering systems, including text-based question answering and ontology-based query answering approaches, as well as hybrid approaches.
- Propose a formalism allowing to represent what could be solved by the ontological reasoning and what will be dedicated to text processing
- Evaluations over specific domain resources including biomedical ontologies and patient health records. An open domain data such as Freebase and DBpedia can be also explored.

Required profile

- Master 2 in Computer Science or related domain
- Knowledge in Semantic Web, Information Extraction, and/or Artificial Intelligence is required. Background in Natural Language Processing, Information Retrieval, or Automatic Reasoning is desired.
- Programming: Java, python
- Language: good English level
- Ability to work in team, motivation on multidiscipline studies

Your work will be rewarded by a gratification according to French laws.

Documents required for application

- CV and motivation letter
- Transcripts for Master and undergraduate courses

Please send applications to yue.ma@lri.fr

References

[1] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications.* Cambridge University Press, 2 edition, 2007.

[2] S. Decker, M. Erdmann, D. Fensel, and R. Studer. Ontobroker: Ontology based access to distributed and semi-structured information. In R. Meersman, Z. Tari, and S. M. Stevens, editors, *Proc. of the 8th Working Conf. on Database Semantics (DS-8)*, volume 138 of *IFIP Conference Proceedings*, pages 351–369. Kluwer, Jan. 1999.

[3] D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

[4] B. Grau. Finding answers to questions, in text collections or web, in open domain or specialty domains. In C. Jouis, I. Biskri, J.-G. Ganascia, and M. Roux, editors, *Next Generation Search Engines: Advanced Models for Information Retrieval*, pages 344–370. Information Science Reference, 2012.

[5] Y. Ma and F. Distel. Concept adjustment for description logics. In *Proceedings of the 7th International Conference on Knowledge Capture (K-CAP 2013)*, pages 65–72, 2013.

[6] A.-L. Minard, A.-L. Ligozat, A. B. Abacha, D. Bernhard, B. Cartoni, L. Deléger, B. Grau, S. Rosset, P. Zweigenbaum, and C. Grouin. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. *JAMIA*, pages 588–593, 2011.

[7] A. Petrova, Y. Ma, G. Tsatsaronis, M. Kissa, F. Distel, F. Baader, and M. Schroeder. Formalizing biomedical concepts from textual definition. *Journal of Biomedical Semantics*, 6(22), 2015.

[8] C. Unger, A. Freitas, and P. Cimiano. An introduction to question answering over linked data. In M. Koubarakis, G. B. Stamou, G. Stoilos, I. Horrocks, P. G. Kolaitis, G. Lausen, and G. Weikum, editors, *Reasoning Web. Reasoning on the Web in the Big Data Era - 10th International Summer School 2014, Athens, Greece, September 8-13, 2014. Proceedings*, volume 8714 of *Lecture Notes in Computer Science*, pages 100–140. Springer, 2014.