Towards Shared Hypothesis Testing in the Biomedical Domain

Asan Agibetov[1], Ernesto Jiménez-Ruiz[2], Alessandro Solimando[3], Giovanna Guerrini[3], Giuseppe Patanè[1], and Michela Spagnuolo[1]

[1] National Council of Research, Genova, Italy,
{asan,giuseppe,michela}@ge.imati.cnr.it
[2] University of Oxford, Oxford, UK, ernesto@cs.ox.ac.uk
[3] University of Genova, Genova, Italy,
{alessandro.solimando,giovanna.guerrini}@unige.it

Abstract. Evidence-based hypothesis testing assumes the existence of a causal chain between the facts. By studying the propagation of evidenced facts in the causal chain (hypothesis) we gain new insights on the progression of a disease. In practice, a hypothesis cannot always be substantiated with a complete asserted knowledge (inability to collect the required evidence), yet it is possible to test a hypothesis with missing knowledge with a lower confidence. In this work we propose a method to perform evidence-based hypothesis testing in the biomedical domain, such that specialists can evaluate confidence of their hypothesis and communicate their findings. We assume that a hypothesis is formalized in an OWL 2 EL ontology and the KB contains incomplete asserted knowledge (ABox). We extract a causal chain from an ontology and represent it as a DAG (node - fact, arc - causal relationship). Users assign importance weights to the facts which they think are more important to support the hypothesis. Evaluation of the hypothesis confidence is then done by computing a weighted sum of fact confidences over the directed path in the DAG (corresponding to the causal chain).

Keywords: biomedical ontology, hypothesis testing, incomplete knowledge

1 Introduction

Some pathologies, such as osteoarthritis (OA), may be evidenced across multiple biological scales (*e.g.*, cellular, molecular, organic, behavioral). An evidence at each biological scale is obtained through the analysis of the results of a specific assay (*e.g.*, cell viability, mechanical properties, MRI, gait analysis). Positive correlation between these evidenced facts is deduced from the statistical experiments or from the literature sources. The positive correlation may be represented as a causal chain of facts (*e.g.*, $f_1 \mapsto f_2 \mapsto \ldots \mapsto f_n$). For such a causal chain to hold – and thus for a hypothesis to be satisfied – every node of this network must be evidenced by a corresponding analysis of an assay. To better convey the idea of hypothesis testing we will focus on the knee articular cartilage degradation due to OA to present our use-case scenario. However the analysis of the causal chain of the evidenced facts is applicable to other case studies.

Use-case. A molecular biologist is studying the cartilage degradation due to OA. The death of chondrocytes (factor f_1) is a common feature of an osteoarthritic cartilage, one way to evidence it is to run cell viability assays, which produce images. By analysing these images, a molecular biologist establishes that there was a decline of cell viability (evidence e_1) [13] (*i.e.*, f_1 is evidenced by e_1). From the literature it was hypothesized that there might be a connection between the death of chondrocytes and joint stiffness (f_2), which affects the gait pattern (e_2) [2]. To support this hypothesis, he collaborates with the Orthopaedics department and obtains an evidence of gait pattern alteration presumably due to cartilage degradation causing joint stiffness. Based on the results he establishes that $\texttt{causes}(f_1, f_2)$ holds, which we denote $f_1 \mapsto f_2$. Both may consider their work done, however there was a jump from *cellular* biological scale to *behavior* biological scale. Since cartilage degradation leads to *Cartilage thinning* (f_3) (surface diminution) it must be seen on *organ* level via analysis of MRI (e_3) [14]. It is therefore possible to refine the hypothesis by adding a new causal relationship, which would mean that instead of $f_1 \mapsto f_2$, we actually need to prove $f_1 \mapsto f_3 \mapsto f_2$.

Hypothesis Testing with Incomplete Knowledge. In practice, the process of data collection and their analysis is very time-consuming and sometimes it is sufficient to have evidences for *some* nodes of the causal chain and not all of them. Alternatively, some of the nodes may be evidenced independently by different research groups and, if combined, they could support a pathology hypothesis. Lastly, some of the causal relationships between the facts may be unsatisfied by statistical results, however they could still be considered due to errors and low confidence in the statistical results. Therefore, it is a common case that a hypothesis is tested with incomplete knowledge.

In this work we propose a method to perform hypothesis testing, assuming that a hypothesis is formalized in an OWL 2 EL ontology with an open-world assumption. We extract a causal chain from an ontology and represent it as a DAG, where each node corresponds to a fact and an edge represents a causal relationship. Users assign importance weights to the facts which they think are more important to support the hypothesis. Evaluation of confidence of a hypothesis is then done by computing a weighted sum of the directed path in the DAG (corresponding to the causal chain).

2 Related Work

Application of Semantic Web technologies in the biomedical domain to infer missing information or new insights in the presence of incomplete knowledge are studied in [10, 12]. In [10] a method for rule-based reasoning with a multi-scale neuroanatomical ontology is presented. Authors conclude that OWL is an important technology for merging disparate data and performing multi-scale reasoning. They demonstrate how OWL-based ontologies and rule-based reasoning help infer novel facts about brain connectivity at large scale from the existence of synapses at a micro scale. Oberkampf *et al.* [12] propose a methodology for interpreting patient clinical data (medical images and reports), semantically annotated by concepts from large medical ontologies. They introduce an ontology containing lymphoma-related diseases and symptoms as well as their relations and use it to infer likely diseases of patients, based on annotations.

In [1, 11] the multi-scale biomedical factors causing cartilage degradation during the OA are considered in a framework for semantic biomedical data exploration. The causal chain is formalized in the *MultiScaleHumanOntology* [6].

Theoretical frameworks to marry formal methods (*e.g.*, First-Order Logic) and probabilistic models (*e.g.*, stochastic processes) are proposed in [15, 9]. In [5], the Stochastic Process Algebra language PEPA [7] was tuned to model biological pathways and complex biological networks, involving stochastic processes. Our work tries to bridge "uncertainty" and "formal methods", similarly to the above methods, but it stays in OWL with a specific application in the biomedical domain.

3 Methodology

Our methodology for hypothesis testing of a multi-scale pathology relies on the formalization of the causal chain, represented as an OWL ontology. Specifically, we consider OWL 2 EL profile, as its axioms are well suited for many of the biomedical formalizations [8]. We assume that the causal chain is modeled by using the existentially quantified restriction axiom to support the open-world assumption similar to anatomical relation modeling in FMA [16, 4] (*i.e.*, $\texttt{Femur} \sqsubseteq \exists \texttt{constitutional\_part\_of}.\texttt{Thigh}$). For instance, the relation between f_3 and f_2 is formalized in MultiScaleHuman ontology [6] as:

$$\texttt{Cartilage\_thinning} \underbrace{\sqsubseteq}_{\texttt{rdfs:subClassOf}} \overbrace{\exists \texttt{causes}.\texttt{Joint\_stiffness}}^{\text{blank node construction}} \tag{1}$$

DAG Representation of Causal Chain. We build a DAG (Directed Acyclic Graph) representation of the causal chain starting from its formalization in an OWL 2 EL ontology. To build a causal chain we analyse EL [3] axioms corresponding to *facts* of the form $A \sqsubseteq \exists R.\ B$, where A and B are atomic DL concepts. That is, we assume that the ontology is built with *atomic concepts* only in the filler of the *existential restrictions*. The existential restriction is applied to a *transitive* relation R, which models *causality* (*e.g.*, *causes*). In the case of nested existential restrictions $\alpha := A \sqsubseteq \exists R_1.\ (\exists R_2.\ C)$ it must be the case that $R_1 = R_2$ and R_1 is transitive. Then we can connect $A \mapsto C$ with R_1 in the DAG.

We analyse the RDF subgraph encoding such axioms and infer a causal relationship between `Cartilage_thinning` and `Joint_stiffness`, as depicted in Figure 1, to obtain a DAG of causal chains $f_1 \mapsto f_2 \mapsto \ldots \mapsto f_n$. We recursively apply the same strategy in the case of nested restrictions.

Importance of Facts and Identification of Missing Nodes. User asserts instances of evidences to the knowledge base, *i.e.*, $\texttt{Cartilage\_thinning}(e_1)$, if such an evidence is found. We also assume that the user provides *importance measures* for the evidences, which we store as an attribute of a node in the DAG. That is, for a node f_i corresponding to some fact we add attribute *importance: 0.45*. To ease the notation we denote it $if(f_i)$. We check if the user asserted any instance for concepts representing evidences f_i, and if so we add attribute *satisfied: True*, otherwise we add *satisfied: False*. We identify the missing nodes by filtering all nodes in the DAG for *satisfied: False*.

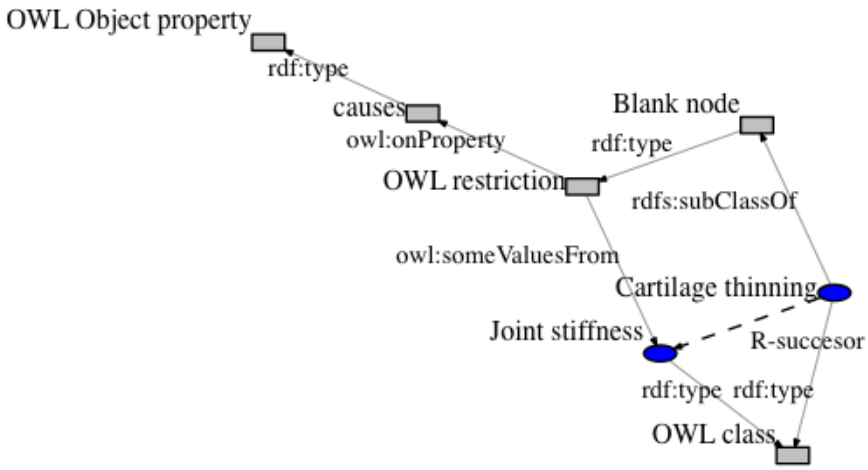

Fig. 1. RDF graph corresponding to causal chain encoding OWL axiom

Assessment of Hypothesis Confidence. To assess the confidence of the hypothesis, we compute the weighted path passing through the facts. For instance, if in our use-case the evidence f_3 was missing, then we would give this node a negative weight (attribute *weight: -1*). By contrast, f_1, f_2 would be given positive weights since we have found an evidence for them. Thus, a weighted path (f_1, f_2) would give us the final confidence value of our hypothesis. For example, $+1 \times if(f_1) + (-1) \times if(f_3) + 1 \times if(f_2)$. We use NetworkX Python library for network analysis tasks (weighted path computation) and basic graph querying (node and arc selection based on attribute values).

4 Discussion and Conclusion

Although the identification of the missing evidenced facts in the hypothesis could have been obtained by executing ad-hoc SPARQL queries, the DAG transformation metaphor allows us to naturally obtain confidence propagation. We consider a rather limited subset of OWL 2 EL axioms during the DAG construction, since the original causal chain formalization, which we built our arguments upon, used only those for causal chain modelling. However, in the future, other types of axioms would be taken into account.

We argue that reasoning on the formalized causal chain may identify the missing nodes which could be presented to the specialist to guide him in the conduction of his experiment, and notify him what is missing for the hypothesis to hold. Alternatively, the knowledge of the missing nodes, necessary to support a hypothesis, may help the specialist in finding a collaboration with other institutions. By using our methodology, depending on the richness of the knowledge model of a hypothesis and importance weights for the facts, researchers may evaluate the confidence of their hypotheses. Moreover, they can also identify what is missing for their hypothesis to be *positive*. It is also possible to publish the results by referring to a knowledge model of causality expressed in a standard ontology language (*i.e.*, OWL).

Acknowledgements

This work was partially funded by the EU Marie Curie ITN MultiScaleHuman (FP7-PEOPLE-2011-ITN, Grant agreement no.: 289897), the EU project Optique (FP7-ICT-318338), and the EPSRC projects MaSI^3, Score!, and DBOnto.

References

1. Agibetov, A., Vaquero, R.M.M., Friese, K.I., Patanè, G., Spagnuolo, M., Wolter, F.E.: Integrated Visualization and Analysis of a Multi-scale Biomedical Knowledge Space. In: EuroVis Workshop on Visual Analytics. pp. 25–29. The Eurographics Association (2014)
2. Andriacchi, T.P., Mündermann, A., Smith, R.L., Alexander, E.J., Dyrby, C.O., Koo, S.: A framework for the in vivo pathomechanics of osteoarthritis at the knee. Annals of Biomedical Engineering 32(3), 447–457 (Mar 2004)
3. Baader, F., Brand, S., Lutz, C.: Pushing the EL envelope. In: In Proc. of IJCAI 2005. pp. 364–369. Morgan-Kaufmann Publishers (2005)
4. Boecker, M.: Teaching Good Biomedical Ontology Design. In: Proceedings ICBO (2012)
5. Ciocchetta, F., Hillston, J.: Bio-PEPA: An Extension of the Process Algebra PEPA for Biochemical Networks. Electronic Notes in Theoretical Computer Science 194(3), 103–117 (Jan 2008), http://www.sciencedirect.com/science/article/pii/S1571066108000285
6. FP7 MultiScaleHuman: MSH Ontology: deliverable reports D8.2 (m24, m36) and OWL file (2015), (accessed July 28, 2015): http://multiscalehuman.miralab.ch/repository/Public_download/D8.2_MSD-Ontology/
7. Hillston, J.: Process algebras for quantitative analysis. In: 20th Annual IEEE Symposium on Logic in Computer Science, 2005. LICS 2005. Proceedings. pp. 239–248 (Jun 2005)
8. Kazakov, Y., Krötzsch, M., Simančík, F.: The Incredible ELK. Journal of Automated Reasoning 53(1), 1–61 (2014), http://dx.doi.org/10.1007/s10817-013-9296-3
9. Kimmig, A., Bach, S.H., Broecheler, M., Huang, B., Getoor, L.: A Short Introduction to Probabilistic Soft Logic. In: NIPS Workshop on Probabilistic Programming: Foundations and Applications (2012)
10. Larson, S.D., Martone, M.E.: Rule-Based Reasoning With A Multi-Scale Neuroanatomical Ontology. In: OWLED (2007)
11. Millán Vaquero, R.M., Agibetov, A., Rzepecki, J., Ondrésik, M., Vais, A., Oliveira, J.M., Patanè, G., Friese, K.I., Reis, R.L., Spagnuolo, M., Wolter, F.E.: A semantically adaptable integrated visualization and natural exploration of multi-scale biomedical data. In: Proceedings MediVis (2015)
12. Oberkampf, H., Zillner, S., Bauer, B.: Interpreting Patient Data using Medical Background Knowledge. In: Cornet, R., Stevens, R. (eds.) ICBO. CEUR Workshop Proceedings, vol. 897. CEUR-WS.org (2012)
13. Ondrésik, M., Correia, C., Sousa, R., Oliveira, J., Reis, R.: Understanding cellular behaviour in early and late stage of MSD. Journal of Tissue Engineering and Regenerative Medicine 8, 412–412 (2014)
14. Pitikakis, M., Chincisan, A., Magnenat-Thalmann, N., Cesario, L., Parascandolo, P., Vosilla, L., Viano, G.: Automatic measurement and visualization of focal femoral cartilage thickness in stress-based regions of interest using three-dimensional knee models. IJCARS (Jul 2015)
15. Richardson, M., Domingos, P.: Markov Logic Networks. Mach. Learn. 62(1-2), 107–136 (Feb 2006), http://dx.doi.org/10.1007/s10994-006-5833-1
16. Rosse, C., Mejino, J.L.V.: A reference ontology for biomedical informatics: the foundational model of anatomy. J. of Biomedical Informatics 36, 500 (2003)