Knowledge Graphs in support of Human-Machine intelligence

Christophe GUÉRET [a,1]
[a] *Accenture Labs, The Dock, Dublin, Ireland*

Abstract. Knowledge graphs are established as a powerful way to represent heterogeneous knowledge in a unified model, providing an integrated view over data coming from different domains. They are the back-end of data fabrics helping businesses make sense of their data and they enable complex, holistic, reasoning via graph machine learning. However the creation of a KG and its use are typically approached as two disconnected tasks even if the use-cases informs the design of an ontology. In this position paper we highlight some of the challenges created by this split view approach. We further discuss how this impairs working towards Hybrid Intelligence and propose an alternative architecture based on several operational components. We argue this set of components can together enable the interactive exchange of ideas between human and machines around a shared data model.

Keywords. HITL, Knowledge Graph

Introduction

With the progress of Artificial Intelligence (AI) came the discussion about whether some day it will replace all human activities. It is now becoming apparent that even if AI has its own strength Human have their own too. We are heading more towards a future combining each other strength in order to achieve a common goal [1,2]. The point can further be made that machines aiming at making sense of human specific traits, such as the appreciation of aesthetics, require crucially depends on the presence of humans in the loop to work [3].

In this context, Knowledge Graphs emerge as a flexible data representation technology capable of connecting together heterogeneous data in a unified semantically rich model [4]. This graph can further be directly used by machine learning approaches to support reasoning [5]. However we observe that the work done around Knowledge Graph typically considers the tasks of creating and putting a KG into use as two separate focus points. We further argue that in order to support creating a Collaborative, Adaptive, Responsible and Explainable AI research agenda [1] there is a need for a tighter integration between components in charge of the usage of the graph and those related to its evolution. Our main contributions are:

- A list of shortcomings and challenges when creating Knowledge Graphs and Use-cases on top of them for a human-machine collective intelligence case;

[1]Corresponding Author: Christophe Guéret christophe.gueret@accenture.com

- An example use-case to ground our discussion;
- A proposal for a high level architecture for a different kind of KG pipeline enabling more collaboration.

For the remainder of this paper we introduce the use-case in Section 2 and the specific challenges this raises (Section 3). We finally conclude on future work after having presented a high level architecture in Section 4.

1. KGs in support of Human-Machine collaboration

Knowledge Graphs are a flexible data representation technology capable of connecting together heterogeneous data in a unified semantically rich model [4]. This graph can be used directly to provide a ”360 view” over data silos in a business context, or constitute the data layer for machine learning algorithms - thereby replacing feature tables used by other machine learning techniques [5]. Several architectures and approaches have been proposed to construct and care for this knowledge graph [6]. Some of the latest advances have recently been discussed in a dedicated workshop [7] which covered, among other topics, the application of graph embedding techniques to support the creation and curation of the graph. This kind of tooling could be of use to Knowledge Scientists [8] in their goal of delivering clean integrated data for use by a set of stake holders.

Next to this, several architectures are being considered to make use of that graph in combination with other forms of machine learning. The success of connectionist models over the past couple of years lead to a particular focus on Neurosymbolic computing combining these models with Knowledge Graphs but symbolic-subsymbolic integration is a larger topic [9]. The graph embedding approach mentioned above is one of those which could be approached via Neurosymbolic computing. A neural network is used to turn the graph into a vector space which is then used for reasoning. This enables usage going beyond querying the graph and reasoning over it with rules.

Connecting the two, the usage of knowledge graphs will see a number of human interventions to design the content of the graph and its usage. The evolution of the graph is also a matter of human-in-the-loop processes [10,8,11] to progressively enrich it along with the consideration of new usages and the tuning of existing ones. Figure 1 is an attempt at representing this interplay between these components and stake holders.

According to Dellermann *et al.* [2] a Hybrid Intelligence usage scenario would exhibit some key Collaboration, Superior Results and Continuous Learning behaviours. This means a greater amount of joint work between the Human and the Machine to tackle a challenge together and with better results than when working alone. Mutual teaching is also key as the Machine need to benefit from the collaboration as the Human will. However the Human-In-The-Loop (HITL) aspects of the approach depicted in Figure 1 may hinder this as this learning will be mediated by another Human. Beyond collaboration, Akata *et al.* [1] add Explainability, Responsibility and Adaptability on the research agenda for Hybrid Intelligence. We in particular highlight here the potential limitation these HITL aspects bring to adaptation. The system designed following an approach as depicted in Figure 1 will have difficulties adapting to new needs from the user without Humans mediating this adaptation.

We propose to consider in a more integrated way the creation and the usage of a Knowledge Graph in order to support a Hybrid Intelligence. To further support this po-

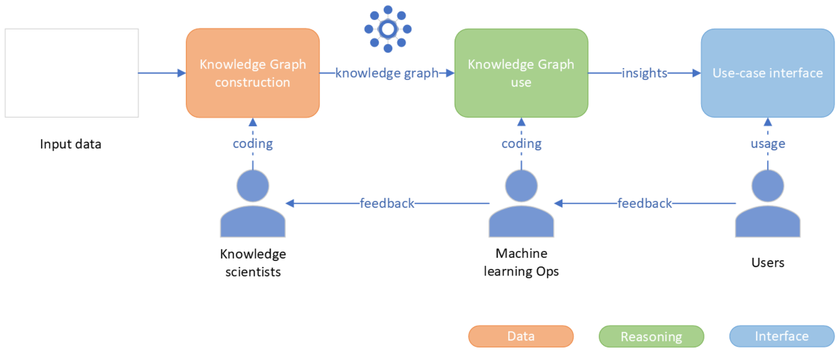

Figure 1. State of the art approaches to making and using a Knowledge Graph in a AI context rely on two distinctive parts besides the use-case interface. Human are in the loop and collectively can improve on the overall, end-to-end, pipeline.

sition we introduce a specific use-case scenario of collaborative diagnosis for the medical domain. The proposed approach is however not limited to health care, we are in fact deploying the same building blocks to support consulting activities around the topic of Industry X.0.

2. Use-case: Consultation Support

We will illustrate this position with an applied use-case in healthcare. In this domain, AI, mainly thank to its machine learning capabilities, is poised to revolutionise our understanding of the human body and support a range of medical processes with valuable insights. Knowledge Graph can in particular support the process of creating multi-omics data sets. This term denotes the combination of several population level data sets (*e.g.* DisGenet [2], PubChem [3]) together with patient level data sets (*e.g.* Synthea[4]) in a unified data model. As all these data sets have a distinct focus (organs, genes, proteins, ...) their combination unlocks a more global understanding of the human body and potentially contributes to a better diagnosis. This kind of graph has been successfully applied, for example, to cluster patients based on a shared medical history contextualised with other data sets [12].

For our use-case we will focus on diagnosis as studied by Richens *et al.* [13]. They formally define the problem as "*The identification of the diseases that are most likely to be causing the patient's symptoms, given their medical history*". Whereas their work focus on using causal reasoning to tackle this challenge we will consider a link prediction task with a graph embedding model. Furthermore, placing ourselves in a Hybrid Intelligence perspective, we will consider the link prediction task being equally performed by Human and Machine operators. Both parties will be expected to contribute to the discussion with justified propositions. They will also be considered to each have a specific expertise complementing that of the other actor, and creating a collective opportunity to learn from each other.

[2] https://www.disgenet.org/, Last visited March 9th 2022
[3] https://pubchem.ncbi.nlm.nih.gov/, Last visited March 9th 2022
[4] https://synthetichealth.github.io/synthea/, Last visited March 9th 2022

Considering this definition Figure 2 represents an overview of main components for this use-case. We will consider a simplified view of one human doctor working with one machine, each having their own knowledge and expertise depicted as a Knowledge Graph. Multi-omics raw data is provided as an input to the human+machine team which collaboratively produces a set of Disease-Symptom links as an output. During this process we aim at seeing the human and the machine debate associations and learn from each other.

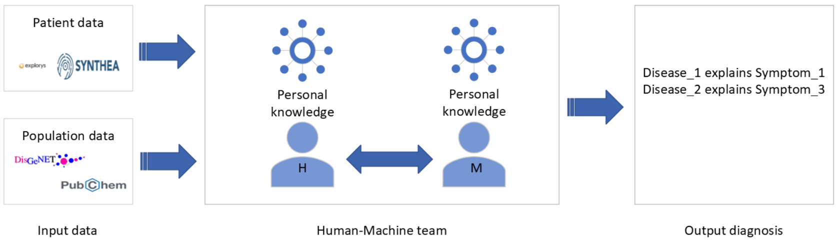

Figure 2. The consultation support use-case generates Disease-Symptom as an output, leveraging multi-omics data as an input. In the middle we find a team made of one human and one machine collaborate to compose the set of links. While doing so they each leverage and improve a personal knowledge base summing up their own expertise and experience. In order to highlight the equal role of both parties the same icon is used with an "H" annotation for the Human intelligence and a "M" for its Machine team mate

Both the human and machine parties are expected to contribute to the task to an equal level and exhibit a shared set of behaviour. These expectations can be summarised as follows:

- Contribute to the refinement of the content of the hypothesis set;
- Learn from the other actor in order to become better at the task over time;
- Explain and argument the propositions made;
- Be mindful of the conclusions made to avoid disclosing potentially sensitive data.

In the next section we go other how these expectation map into the process of creating and using a Knowledge Graph as represented in Figure 1. We in particular highlight the specific challenges we identify from this kind of pipeline.

3. Implementation challenges

We hereafter consider the implementation of the use case introduced in Section 2 in the light of the design approach depicted in Figure 1.

The problem consists in assembling a multi-omics dataset and executing a link prediction task. Considering an ontology introducing the specific concepts of Disease and Symptom that prediction task will consist in linking a set of identified Symptom nodes to a set of Disease nodes. This can be implemented using tools such as the Enterprise Knowledge Graph platform Stardog [5] and the graph machine learning library Ampligraph [14]. Explanations for the links could be provided as information about the importance of specific other edges in the graph for the prediction [15].

[5] https://www.stardog.com/, Last visited March 11th 2022

Paraphrasing the research agenda for Hybrid Intelligence system proposed in [1] we can define 4 set of target behaviour for the assembled system:

- Collaboration: the system works in synergy with humans;
- Adaptation: the system adapt to the humans and their environment;
- Responsibility: the system will behave ethically and responsibly;
- Explanability: the system will share its awareness, goals and strategy.

We can then study for each of those 4 points how an approach based on constructing a graph, shipping it to a link prediction module, and providing users with a ranked list of suggestion could be challenged by these objectives.

Collaboration requires providing suggestion for links as well as receiving some. Instead of being a one-off output the link prediction task will have to be placed in an iterative context of way and back propositions between different parties discussing it. There is thus a need for a feedback loop between reasoning and the validation of each hypothesis, or subset thereof.

Adaptation to the work context requires being able to contextualise the graph to a particular patient. Humans will be expected to switch from working one patient case to another one and the machine should adapt to that. There are also parts of the discussion which may require focusing on sub-parts of the graph and require more external data. For instance, validating a working hypothesis against the most recent research published on BioRxiv[6]. It can also be expected from the machine to adapt its understanding of the medical landscape over time as it build up expertise. This potentially means refining the ontology and/or the data over the course of the interactions. A process akin to reinforcement learning.

Responsibility towards the data and the insights provided relates to other challenges in the creation of central Knowledge Graphs containing potentially sensitive information [16]. The AI will have to be mindful of not disclosing potentially sensitive information via its link predication or the explanation for them. On the other hand, as it will collaborate with and adapt to other parties the AI should not take for granted everything said by everyone. Each piece of incoming information will have to be parked and scrutinized before it makes it into further reasoning process.

Explanability will be key to the collaboration in order for all the actors to learn from each other. This means the machine will have to explain its reasoning for each single link proposed [17] but also for a set of links as a whole; potentially with a Global Counterfactual Explanation (GCE) [18] approach. Conversely, humans will do the same and explain their train of thought. That knowledge should then be considered by the machine for its own further reasoning approach (with a pinch of salt, as discussed above).

From the above challenges we identify three main aspects a state of the art solution would miss: the interactive construction and explanation around a set of link predictions; the acquisition of feedback from the humans about facts and reasoning processes; and, finally, the progressive validation of this external knowledge based on consensus. In the next section we revisit Figure 1 and propose with Figure 3 an approach to position these missing blocks.

[6] `https://www.biorxiv.org/`, Last accessed March 11th, 2022

4. Proposal for an hybrid approach

As outlined above, our target use-case of building an AI working together with Human doctor to refine a set of diagnosis hypothesis will face several challenges. We argued that state of the art approach would miss the target on some aspects and highlighted some major missing components. In Figure 3 we propose a more detailed version of Figure 1 enhanced with those missing building blocks. We also replaced the emphasis on HITL processes by an emphasis on feedback loops within the system itself.

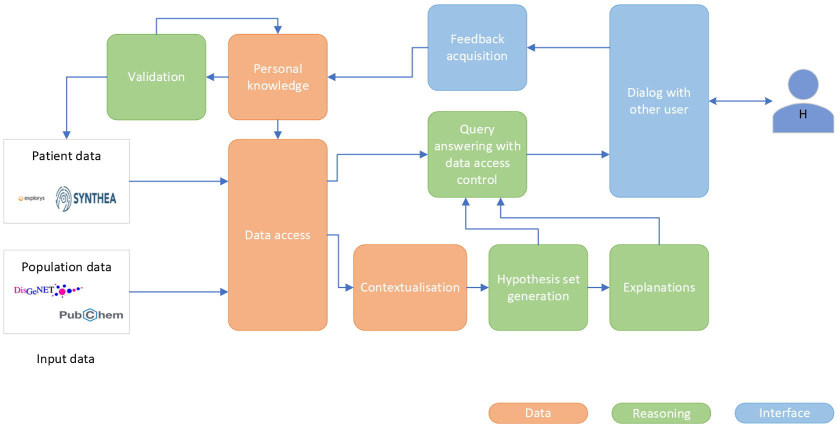

Figure 3. Overview of the different functional components of the proposed end-to-end pipeline. This is here depicted from the point of view of the machine having a dialog with the Human but we can expect the alternative view to follow the same logical process. What is worth highlighting in this diagram is the presence of feedback loop and components dedicated to supporting, and benefiting from, the interaction with the human.

The different components serve different purposes along the same topics highlighted in Figure 1:

Data is managed via a data access layer mapping the raw data into the Knowledge Graph. This mapping is informed by the ontology which is part of the personal knowledge of the agent. With this latter component the agent will also keep track of additional data not found in the input data. All the data goes through a contextualisation layer which sub-sets it to focus on a particular part most relevant to the current interaction with the Human user;

Reasoning over the data means giving access to the input data via answering queries, and reasoning over the potential risks of such disclosure. It also means giving access to the link prediction task output is provided together with the explanations for it. The last aspect of reasoning is the validation which will act like a consensus reasoner taking a decision about trusting the external data gathered over time. These decisions may result in new ground truth for the personal knowledge or new facts to return to the input data as globally true facts.

Interface with the user is around sharing data and insights as well as gathering feedback. This feedback is impacting the content of the personal knowledge by means of additional data and/or refinement of the ontology. Tentatively at first and more systematically later once the feedback has been validated.

As outlined earlier, Figure 3 keeps this focus on the use-case of consultation support, with the aim of working together with human medical doctors to refine a set of diagnosis hypothesis. But we are currently researching and implementing the same pipeline for an Industry X.0 [19] context. In this non healthcare related context the problem is however similar: consultants are tasked with figuring out which kind of solution set could tackle a particular challenge. Industry X.0 is a broad concept covering several aspects of the digital transformation of industries. Finding the best association set between solutions and challenges is a matter of a collective effort leveraging different opinion and expertise.

Some parts of this proposed pipeline have already been put in production for business solutions [16] and some others have been more formally identified [20,21]

Conclusion

The future of AI is a collaboration between Human and Machine in order to leverage each other intelligence's specific strength. In this context Knowledge Graphs can play a key role as a data integration tool and data model for reasoning We argue that however approaches such as ontology based data integration and link predication each tackle one of those two aspect as a distinct focus (creation versus usage of a KG) and propose to look for opportunities to connect them closer. Furthermore we note that in order to be adaptive and AI intelligence working in a hybrid context should have a focus on gathering feedback from humans as well as fellow machines, and reason over this feedback. Our proposal is a high level architecture made of Data, Reasoning and Feedback focused technical components addressing this opportunity. We more specifically instantiate this architecture on a particular task of hypothesis set curation and report on early implementation and research work done for this architecture.

References

[1] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K.V. Hindriks, H.H. Hoos, H. Hung, C.M. Jonker, C. Monz, M.A. Neerincx, F.A. Oliehoek, H. Prakken, S. Schlobach, L.C. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen and M. Welling, A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence, *Computer* **53**(8) (2020), 18–28. doi:10.1109/MC.2020.2996587.

[2] D. Dellermann, P. Ebel, M. Soellner and J.M. Leimeister, Hybrid Intelligence (2021). doi:10.1007/s12599-019-00595-2.

[3] H.J. Wilson and P.R. Daugherty, Robots Need Us More Than We Need Them, *Harvard Business Review* (2022). https://hbr.org/2022/03/robots-need-us-more-than-we-need-them.

[4] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J.E.L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.-C.N. Ngomo, S.M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab and A. Zimmermann, Knowledge Graphs (2020). doi:10.1145/3447772. https://arxiv.org/abs/2003.02320.

[5] X. Wilcke, P. Bloem and V.D. Boer, The knowledge graph as the default data model for machine learning (2017), 1–13. https://datasciencehub.net/paper/knowledge-graph-default-data-model-learning-heterogeneous-knowledge.

[6] S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P.N. Mendes, B.V. Nuffelen, C. Stadler, S. Tramp and H. Williams, Managing the Life-Cycle of Linked Data with the LOD2 Stack, in: *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part II*, P. Cudré-Mauroux, J. Heflin,

E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J.X. Parreira, J. Hendler, G. Schreiber, A. Bernstein and E. Blomqvist, eds, Lecture Notes in Computer Science, Vol. 7650, Springer, 2012, pp. 1–16. doi:10.1007/978-3-642-35173-0_1.

[7] D. Chaves-Fraga, A. Dimou, P. Heyvaert, F. Priyatna and J.F. Sequeda (eds), Proceedings of the 2nd International Workshop on Knowledge Graph Construction co-located with 18th Extended Semantic Web Conference (ESWC 2021), Online, June 6, 2021, in *CEUR Workshop Proceedings*, Vol. 2873, CEUR-WS.org, 2021. http://ceur-ws.org/Vol-2873.

[8] G. Fletcher, P. Groth and J.F. Sequeda, Knowledge Scientists: Unlocking the data-driven organization, *CoRR* **abs/2004.07917** (2020). https://arxiv.org/abs/2004.07917.

[9] F. van Harmelen and A. ten Teije, A Boxology of Design Patterns for Hybrid Learning and Reasoning Systems (2019). doi:10.13052/jwe1540-9589.18133.

[10] J.F. Sequeda, W.J. Briggs, D.P. Miranker and W.P. Heideman, A Pay-as-you-go Methodology to Design and Build Enterprise Knowledge Graphs from Relational Databases, in: *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I.F. Cruz, A. Hogan, J. Song, M. Lefrançois and F. Gandon, eds, Lecture Notes in Computer Science, Vol. 11779, Springer, 2019, pp. 526–545. doi:10.1007/978-3-030-30796-7_32.

[11] PoolParty, An Enterprise Knowledge Graph Life Cycle - A Summary - PoolParty Manual 8.1. https://conf-doc.poolparty.biz/pp/white-papers-release-notes/poolparty-technical-white-paper/an-enterprise-knowledge-graph-life-cycle-a-summary.

[12] D. Cahalane, A. Kouznetsov and C. Guéret, Knowledge Graphs to help with Data-driven Clinical Decision-making, in: *Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24-28, 2021*, O. Seneviratne, C. Pesquita, J. Sequeda and L. Etcheverry, eds, CEUR Workshop Proceedings, Vol. 2980, CEUR-WS.org, 2021. http://ceur-ws.org/Vol-2980/paperx.pdf.

[13] J.G. Richens, C.M. Lee and S. Johri, Counterfactual diagnosis (2019). doi:10.48550/arxiv.1910.06772. http://arxiv.org/abs/1910.06772.

[14] L. Costabello, S. Pai, C.L. Van, R. McGrath, N. McCarthy and P. Tabacof, AmpliGraph: a Library for Representation Learning on Knowledge Graphs, 2019. doi:10.5281/zenodo.2595043.

[15] M.T. Keane and E.M. Kenny, How Case-Based Reasoning Explains Neural Networks: A Theoretical Analysis of XAI Using Post-Hoc Explanation-by-Example from a Survey of ANN-CBR Twin-Systems, *Lecture Notes in Computer Science* (2019), 155–171–. ISBN 9783030292492. doi:10.1007/978-3-030-29249-2_11.

[16] M. Kujawinski, C. Guéret, C. Kumar, B. Woods, P. Klinov and E. Sirin, On Constructing Enterprise Knowledge Graphs Under Quality and Availability Constraints, in: *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24-28, 2021, Proceedings*, A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P.M. Barnaghi, A. Haller, M. Dragoni and H. Alani, eds, Lecture Notes in Computer Science, Vol. 12922, Springer, 2021, pp. 699–713. doi:10.1007/978-3-030-88361-4_41.

[17] F. Bianchi, G. Rossiello, L. Costabello, M. Palmonari and P. Minervini, Knowledge Graph Embeddings and Explainable AI, in: *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, IOS Press, 2020, pp. 49–72.

[18] G. Plumb, J. Terhorst, S. Sankararaman and A.S. Talwalkar, Explaining Groups of Points in Low-Dimensional Representations, *Proceedings of machine learning research* **119** (2020), 7762–7771.

[19] E. Schaeffer, *INDUSTRY X.0, Realizing digital value in industrial sectors*, 2022. https://www.accenture.com/sg-en/insight-realizing-digital-value-industrial.

[20] C. Gueret and L. Costabello, Systems and methods providing evolutionary generation of embeddings for predicting links in knowledge graphs, 2021, US Patent App. 16/590,738.

[21] F. Lecue, C.D.M. Gueret and D.J. Cahalane, Knowledge graph weighting during chatbot sessions, 2021, US Patent 11,126,919.