Dark Knowledge: Knowledge Graphs of What Cannot Happen or Should not Be Done *

Oleg Sychev[1][0000−0002−7296−2538]

Volgograd State Technical University, Lenin Ave, 28, Volgograd, 400005, Russia
oasychev@gmail.com

Abstract. While a significant number of large knowledge graphs is developed during the last years, they are mostly used for information search. The ability to reason conclusions on them is limited, and a lot of works turn to approximate methods like neural networks and graph embeddings to draw conclusions and use the accumulated knowledge. In this work, I argue that in human thinking the main usage of logical reasoning is not making far-fetched strict logical conclusions but weeding out obviously wrong ideas and false information while generating new ideas is mostly intuitive. This can be used in constructing hybrid reasoning systems where neural networks play the role of intuition (generating possible solutions) while logical reasoning is used to verify and filter these solutions. This requires creating knowledge graphs containing negative information: the information of what cannot happen or should not be done and why. The problems of creating negative knowledge graphs are discussed. Hybrid systems using negative knowledge graphs will be a lot more trusted as they will have a system of human-verifiable rules that guarantees avoiding the worst errors and filter possible solutions which can be used in many fields from decision making to natural-language parsing.

Keywords: Knowledge graphs · Hybrid reasoning.

1 Introduction

During the last years, knowledge graphs of impressive sizes were developed in different subject fields like medicine [12], programming [1], scholarly data [19], and commonsense [13, 20] knowledge. However, these knowledge graphs are mostly focused on simple factual knowledge; they are used for information search and retrieval. The amount of reasoning performed on these knowledge graphs is minimal; it is often limited to pattern search (like program static analysis [16]).

The results of this progress is a large body of knowledge with limited abilities to draw conclusions from it. We have too much recorded knowledge but too little comprehension. While theoretically, the highly-structured data in RDF and OWL should be ideal for precise logical reasoning, practically when it comes to applying the accumulated knowledge, a significant number of researchers use

* The reported study was funded by RFBR, project number 20-07-00764.

frequencies, embeddings, neural networks, and other approximate methods to exploit the accumulated knowledge. These methods have their own disadvantages caused by the high-dimensionality of the input data [5].

Neural networks, embeddings, and statistical methods gained ground in hybrid systems in query answering [8], knowledge-base completion [10], generating natural-language representations of the graph [14], and hybrid reasoning [23]. Many of the hybrid works, aimed at finding new axioms or making conclusions from available data, use knowledge graphs as inputs for neural networks so that neural networks or other statistical methods make final conclusions, unverified (at least in-system) by logical reasoning. This may result in a decent percentage of correct answers as neural networks are a good way of building approximations of the target distribution. However, much less effort is spent on analysing the kind of errors neural networks with graph embeddings do.

One of the common problems of neural networks is their poor stability. Studying adversary examples and other forms of attacks [25, 11, 17] shows that they are easily capable of errors of any magnitude which are uncharacteristic for human thinking, for example classifying indistinguishable for human eye images differently. Generative language models like GPT-2 [18] are good at capturing the topic and the mood of text fragment when generating text completion, but they make serious logical mistakes often and fail to learn some rather simple grammar rules concerning determiners and gender-related pronouns. For example, starting from a sentence with the subject "a boy", it can continue to use "a boy" through the generated text, generating collocations like "a boy's father" and even "a boy's fingers." It also generates sentences like "The girl smiled, and looked to his father" without any clear reference to "his" in the previous sentences. The logical mistakes are more often and obvious, even in short phrases like "The cows ate their own blood" or "They laid on each other." What is more concerning for building a trustworthy AI, GPT-2, given innocent verbs like "kiss," often generates sex scenes without any restrictions on the age and kinship of the participants.

This poses a problem in absence of the ways to weed out erroneous results. The metrics used in current studies on question answering, for example, favour the percent of the correct answers [26] – but they do not take into account the severity of the errors – while in reasoning and planning, making a few small errors is often preferable to making a single big error. Changing the metrics to reflect the degree of being wrong may help improve assessments of question-answering systems, stimulating their development.

Works aimed at analysing and correcting a neural network's output mostly concentrate on using another neural network or statistical methods [24, 7, 6] to achieve this which would have the same problems [21]. Is there a way to create a hybrid system where reasoning and neural networks will interact for creating a better AI? In order to understand this, we need to analyse how human thinking works. How much our reasoning abilities match these of a typical software reasoner?

2 What do humans use reasoning for

Software reasoners either try to infer all possible logical consequences of a set of axioms or work backwards from a set of goals to axioms. While humans are capable of inferring logical conclusions from the available knowledge, making correct, far-reaching conclusions requires special training – and not everyone is capable of that, especially if these conclusions are made using complex formal models like in mathematics and programming. Introducing human learners to formal logic and mathematics is a difficult task with a high dropout rate. It is not something every developed human personality is capable of doing well. This means that making far-reaching, logically-strict conclusions is not what human thinking was developed for – not its primary evolutionary role. What can it be? What human reasoning excels at?

We can exclude tasks that are better done without reasoning. Our brains automate any activity we do routinely, even if it includes fairly high-level cognitive tasks. An experienced lecturer can make a lecture on a known subject without thinking much; singers are known to repeat well-known songs even while being drunk with all intonations.

But there is another way of thinking that also avoids conscious logical reasoning: intuition and creativity. It allows creating new ideas without tracing them logically to the available knowledge, and when a human "wonders" about a complex problem intuitively, it is often internally silent – i.e. it lacks conscious verbal thoughts. By itself, this proves only that intuition works unconsciously; it may still be unconsciously logical.

However, looking at the tasks typically used for studying intuition we can see that these are the tasks where the solution can be easily verified once it is found while iterating all possible solutions is difficult e.g. matchstick arithmetic [9, 15] or anagrams [2]. Intuition excels at the tasks where the solution is hard to find but easy to check. Historical examples show that when used for more complex tasks, intuition can give us "harmonic" (as Poincaré wrote) but wrong answers [4]. So the defining features of intuition are its abilities to iterate or even generate new ideas easily – and perform some kind of optimisation, finding nice candidate solutions. However, these solutions, when studied rigorously, can be wrong. This is similar to what generative artificial neural networks do.

These intuitive candidate solutions must be checked for possible pitfalls and violating commonsense. This is especially important for living organisms, as – unlike, say, a chess-playing program – living beings do not get a second chance after making a lethally-wrong decision. And checking possible errors and pitfalls, assessing the believability of the given idea (according to the existing world model), is something that logical thinking does well. While generating new creative ideas through intuition is often non-traceable, they just "come" into our heads – reality-checking of these ideas is logical and traceable. When humans think that some idea is stupid, they often have a particular reason why it is. So we can conclude that one of the primary functions of human logical reasoning is weeding out wrong and dangerous ideas.

Creating hypotheses and weeding out unbelievable ones is crucial in a wide range of human intelligence tasks, from anaphora resolution and sentence parsing to question answering and assessing believability of texts and ideas.

The current AI models created using reinforced learning mostly lack logical subsystems of validating new strategies, but they lack reasons to develop something like this because of the absence of risk: few reinforced-learning tasks include the risk for the model to die (i.e. stop acting, gathering experience, and not being able to propagate itself) after performing a too-wrong action. This may stop them from developing human-like learning because a reinforced-learning agent learns only the necessary skills. Careful examination of the examples from OpenAI multi-agent autocurricula project [3] shows that the hiders learned to build their "fort" exactly in the time necessary to be safe from the seekers, and while they are capable of sophisticated cooperation if the time restrictions demand it, they also make inefficient moves if the time allows that.

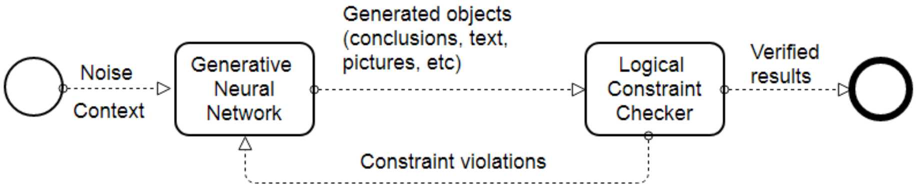
Noise
Context
Generative Neural Network
Generated objects (conclusions, text, pictures, etc)
Logical Constraint Checker
Verified results
Constraint violations

Fig. 1. Hybrid human-like AI system using logical reasoning for assessing "intuitively" generated results.

So, we can roughly see a human-like intelligence as a system where neural networks generate new ideas from given context and random noise (in case the existing ideas are not sufficient to solve the problem at hand), logical reasoning assesses their applicability and the level of risk using available knowledge before trying them, then the ideas that passed logical check are implemented under conscious control and – if successful and repeated many times – get automatised, leaving conscious domain again.

The logical reasoning plays the role of the control contour over intuition to keep the person safe from nicely-looking but grievous errors. This can be a good design for hybrid AI systems as well. Such a system (see Fig. 1) will function like Thaler's Creative Machines but using a reasoner instead of a perceptron. The found constraint violations will become a part of the error function for the neural network so that it will learn to avoid breaking constraints on its own while the people using the hybrid system can maintain the set of logical restrictions, adding new rules as necessary.

3 Negative knowledge graphs for hybrid human-like intelligence

So good possible use of logical reasoning is assessing the results generated by neural networks and weeding out the obviously wrong ones. This allows significantly increasing the resulting system's stability (fewer major changes of the result because of small changes in the input) and trustworthiness (avoiding the worst errors) as such hybrid system, while being creative, will have a human-readable set of limits on the decisions it makes. The percentage of the correct results will be improved at the cost of worst errors which is good for usability because limiting the severity of errors is often more important than limiting their quantity.

However, this poses a unique challenge to the knowledge graphs and ontologies used for reasoning. They need to concentrate on finding errors – i.e. they should contain knowledge about what cannot happen or should not be done. And this is something most modern knowledge graphs miss.

Currently developed knowledge graphs mostly contain positive knowledge – knowledge about what can happen or happened. Under the open-world assumption, this cannot be extended to make conclusions about unbelievable events: if something is missing, it is not conclusive whether it is impossible or it is just an omission. For example, ConceptNet [22] contains the information that a dog is capable of barking, being a pet, and playing frisbee. But it does not show us that a dog is incapable of meowing, keeping a pet, or throwing a frisbee. The absence of these links means nothing as, according to ConceptNet, a dog has no ”capable of” link to ”sit”, and the only link to sleep is to ”sleep a long time” and to jumping is ”jump over a log”. What should an error-checking reasoner conclude from it? Can a dog sleep for a short time or jump over something else than a log? ConceptNet also states that a dog is capable of winning a blue ribbon, but it does not say anything about winning red or green ribbons – or winning other prizes. According to ConceptNet, a dog can please its master and run away from its master, but can a dog make its master angry or run to its master? Nothing could be inferred from ConceptNet about this.

These examples show structural problems of existing knowledge graphs when used for reasoning, assessing believability, and error checking. These graphs lack relevant information. This negative knowledge – the knowledge of what to avoid and what cannot happen – plays a crucial role in human reasoning, but it is largely absent from knowledge graphs, just as dark matter is thought to account for a large percentage of mass in the universe while not been seen. We can call this form of knowledge the dark knowledge.

4 Problems of negative knowledge graphs construction

Developing negative (or dark) knowledge graphs poses new challenges. First, the usual techniques of creating large graphs by mining a large set of natural-text sources work poorly for mining dark knowledge. Most texts are devoted to what happened or could happen; the information on what cannot happen is mostly implicit or absent. Humans are so good at detecting unbelievable ideas that they rarely feel the need to write the necessary knowledge down, much less in a formal form. Crowdsourcing may work better, possibly combined with the gamified approach to gathering commonsense dark knowledge (e.g. a game like "teach an alien about living on Earth.")

The second problem is that the list of actions that someone cannot do – or that cannot be done with something – is far larger than the list of actions that can be done. To store this efficiently, a well-structured taxonomy is necessary so that we would avoid storing separately the facts that dogs, cats, and fish all cannot throw a frisbee. It is necessary to precisely limit the classes that can and cannot perform certain actions (i.e. classes like "does not have hands," – note that under the open-world assumption, not belonging to the class "has hands" means nothing – or "is too heavy for being thrown," etc). These will be necessary when providing explanations of the system's conclusion (e.g. "a dog cannot throw a frisbee because a dog lacks hands"). Precise identification of these borders is not necessary for regular (positive) knowledge graphs under the open-world assumption, but it is necessary for dark-knowledge graphs.

Also, we should take into account different degrees of impossibility. Some things are totally impossible like a human person throwing a building. But some things are possible but should not appear often in the results because they are rare. We can distinguish simply rare events from events requiring special context so that generating texts about dogs pulling a sleigh would be done only for the appropriate breed of dogs while kissing a mother would not lead to generating obscene texts without further indications that this is the case. Enhancing graphs with the information about frequency or likelihood-to-happen for the events is possible, but the preferable approach is identifying the proper context that makes it possible. So negative knowledge graphs and ontologies should contain a fair degree of complex statements, identifying contexts for suitability of different actions. This makes them more complex than regular, positive graphs that mostly contain simple, factual statements.

5 Conclusion

One of the primary roles of logical reasoning in human thinking is assessing the believability of new ideas, both coming from external sources (learning) and intuition (creativity). To become trustworthy, AI systems should employ the same strategy, relying on a clear, human-verifiable set of rules, constraining their output. These systems will be able to guarantee minimum standards of reliability for the generated solution.

To develop such systems, we need to formalise dark knowledge: build knowledge graphs containing the information of what cannot or should not be done. While creating such graphs requires developing new methodologies and overcoming challenges, it opens the path for increasing the reliability – and, implicitly, creativity as wider guesses would be possible – of hybrid AI systems.

References

1. Abdelaziz, I., Dolby, J., McCusker, J.P., Srinivas, K.: Graph4code: A machine interpretable knowledge graph for code (2020), https://arxiv.org/abs/2002.09440
2. Ammalainen, A., Moroshkina, N.: When an error leads to confidence: False insight and feeling of knowing in anagram solving. Psychology. Journal of the Higher School of Economics **16**, 774–783 (12 2019). https://doi.org/10.17323/1813-8918-2019-4-774-783
3. Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., Mordatch, I.: Emergent tool use from multi-agent autocurricula. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SkxpxJBKwS
4. Boden, M.A.: The Creative Mind: Myths and Mechanisms. Basic Books, Inc., USA (1991). https://doi.org/10.5555/102753
5. Donoho, D.: High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture pp. 1–32 (01 2000)
6. Gorban, A.N., Burton, R., Romanenko, I., Tyukin, I.Y.: One-trial correction of legacy ai systems and stochastic separation theorems. Information Sciences **484**, 237–254 (May 2019). https://doi.org/10.1016/j.ins.2019.02.001
7. Gorban, A., Golubkov, A., Grechuk, B., Mirkes, E., Tyukin, I.: Correction of ai systems by linear discriminants: Probabilistic foundations. Information Sciences **466**, 303–322 (Oct 2018). https://doi.org/10.1016/j.ins.2018.07.040
8. Kalo, J.C., Fichtel, L., Ehler, P., Balke, W.T.: Knowlybert - hybrid query processing over language models and knowledge graphs. In: International Semantic Web Conference (ISWC). Springer International Publishing, Athens, Greece (11/2020 2020)
9. Knoblich, G., Ohlsson, S., Haider, H., Rhenius, D.: Constraint relaxation and chunk decomposition in insight problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition **25**, 1534–1555 (11 1999). https://doi.org/10.1037/0278-7393.25.6.1534
10. Kolyvakis, P., Kalousis, A., Kiritsis, D.: Hyperbolic knowledge graph embeddings for knowledge base completion. In: The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12123, pp. 199–214. Springer (2020). https://doi.org/10.1007/978-3-030-49461-2_12
11. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=HJGU3Rodl
12. Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., Sun, Z., Tang, B., Chang, T.H., Wang, S., Liu, Y.: Real-world data medical knowledge graph: construction and applications. Artificial Intelligence in Medicine **103**, 101817 (2020). https://doi.org/https://doi.org/10.1016/j.artmed.2020.101817
13. Liu, H., Singh, P.: Conceptnet — a practical commonsense reasoning tool-kit. BT Technology Journal **22**(4), 211–226 (Oct 2004). https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d, https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d
14. Moussallem, D., Gnaneshwar, D., Castro Ferreira, T., Ngonga Ngomo, A.C.: Nabu - multilingual graph-based neural rdf verbalizer. In: International Semantic Web Conference (ISWC). Springer International Publishing, Athens, Greece (11/2020 2020)

15. Øystein Olav Skaar, Reber, R.: Motivation through insight: the phenomenological correlates of insight and spatial ability tasks. Journal of Cognitive Psychology **0**(0), 1–13 (2020). https://doi.org/10.1080/20445911.2020.1844721
16. Pattipati, D.K., Nasre, R., Puligundla, S.K.: OPAL: An extensible framework for ontology-based program analysis. Software: Practice and Experience **50**(8), 1425–1462 (Mar 2020)
17. Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., Raffel, C.: Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 5231–5240. PMLR, Long Beach, California, USA (09–15 Jun 2019), http://proceedings.mlr.press/v97/qin19a.html
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019), https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
19. Salatino, A., Osborne, F., Motta, E.: Ontology Extraction and Usage in the Scholarly Knowledge Domain, pp. 91–106. IOS Press (11 2020). https://doi.org/10.3233/SSW200037
20. Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N.A., Choi, Y.: Atomic: An atlas of machine commonsense for if-then reasoning. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 3027–3035 (Jul 2019). https://doi.org/10.1609/aaai.v33i01.33013027
21. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D.: Hidden technical debt in machine learning systems. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. p. 2503–2511. NIPS'15, MIT Press, Cambridge, MA, USA (2015). https://doi.org/10.5555/2969442.2969519
22. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. p. 4444–4451. AAAI'17, AAAI Press (2017). https://doi.org/10.5555/3298023.3298212
23. Stoilos, G., Juric, D., Wartak, S., Schulz, C., Khodadadi, M.: Hybrid reasoning over large knowledge bases using on-the-fly knowledge extraction. In: The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12123, pp. 69–85. Springer (2020). https://doi.org/10.1007/978-3-030-49461-2_5
24. Thaler, S.: The creativity machine paradigm: Withstanding the argument from consciousness. APA Newsletters **11**, 19–30 (01 2012)
25. Tyukin, I.Y., Higham, D.J., Gorban, A.N.: On adversarial examples and stealth attacks in artificial intelligence systems. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–6 (2020). https://doi.org/10.1109/IJCNN48605.2020.9207472
26. Usbeck, R., Röder, M., Hoffmann, M., Conrads, F., Huthmann, J., Ngonga Ngomo, A.C., Demmler, C., Unger, C.: Benchmarking question answering systems. Semantic Web **10**, 1–12 (08 2018). https://doi.org/10.3233/SW-180312