Some Insights Towards a Unified Semantic Representation of Explanation for eXplainable Artificial Intelligence (XAI)

Ismaïl Baaj
CEA, LIST
91191 Gif-sur-Yvette cedex, France.
ismail.baaj@cea.fr

Jean-Philippe Poli
CEA, LIST
91191 Gif-sur-Yvette cedex, France.
jean-philippe.poli@cea.fr

Wassila Ouerdane
MICS, CentraleSupélec
Université Paris-Saclay,
Gif sur Yvette, France.
wassila.ouerdane@centralesupelec.fr

Abstract

Among challenges for eXplainable Artificial Intelligence (XAI) is *explanation generation*. In this paper we put the stress on this issue by focusing on a semantic representation of the content of an explanation that could be common to any kind of XAI. We investigate knowledge representations, and discuss the benefits of conceptual graph structures for being a basis to represent explanations in AI.

1 Introduction

Today eXplainable Artificial Intelligence (XAI) is recognized as a major need for future applications. It aims at producing intelligent systems that reinforce the trust of the users (Mencar and Alonso, 2018), who desire to understand automatic decision (Alonso et al., 2017). Moreover, it is part of a context where laws reinforce the right of users (European Council, 2016; US Council, 2018). These last years, many XAI systems have emerged with various applications such as automatic image annotation (Pierrard et al., 2019), recommender systems (Chang et al., 2016) or decision making (Wulf and Bertsch, 2017; Baaj and Poli, 2019).

So far, the researches focus mainly on two specific points. On the one hand, the literature is abundant about the production of the content of the explanation (Biran and Cotton, 2017; Gilpin et al., 2018). On the other hand, different papers focus on the difficult task of evaluation (Mohseni et al., 2018; Hoffman et al., 2018). However, an interesting and not easy question has motivated few works, namely the structure of an explanation (see for instance, (Overton, 2012) for the scientific explanation case).

Despite the several existing XAI approaches, we believe that they all share the need to provide at the end an explanation in natural language. We propose to meet this need through a semantic representation of the content of an explanation. We dedicate this paper to discuss the construction of such a representation by highlighting the different criteria and characteristics that we think this representation should meet to be a unified framework for XAI. Especially, we will discuss a particular representation namely conceptual graphs (Sowa, 2000), and its derivatives, that we believe offer a great potential for this kind of representation.

The paper is organized as follows: in Section 2, we motivate the need of a semantic representation for generating explanations in a XAI architecture. Next, in Section 3, we continue with an overview of some existing knowledge representations in AI, pointing out some of their weaknesses regarding our needs. It leads us to present some narrative representation models in Section 4 and to focus in particular on a semantic network used for text representation. We discuss this one in Section 5, regarding its potential as a semantic representation of explanation in AI. Finally, we conclude with some research perspectives in Section 6.

2 Motivations

We aim in this work to answer the need of providing an explanation in natural language for XAI. To account for this, we propose to abstract the process of generating explanations, as shown in Figure 1. The idea is to represent the explanation generation process through three major components:

- *the content extraction* from an instantiated AI model,
- *the semantic representation* of this content, and
- *the text generation* by relying on Natural Language Generation (NLG).

The content extraction is specific to each model (e.g. decision trees, expert systems, etc.): it takes

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 14–19,
Tokyo, Japan, October 29 2019.

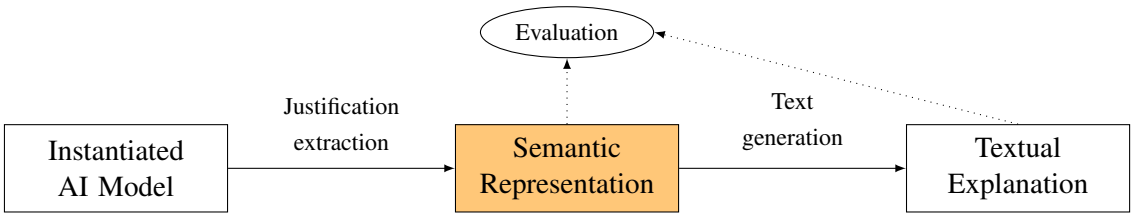

Figure 1: XAI architecture proposal to produce and evaluate explanations

as input the instantiated model, i.e. all the internal values of the model for a given input: for instance, a neural network and the values of all the weights, the execution trace of an expert system, etc. On the contrary, the other components are common to all kind of models and the research efforts can though be factorized. The generation of text from a semantic representation can be helpful for multilingual support. This split may also help the evaluation phase by allowing to separate the target of evaluations to independent steps: e.g., the content of the explanation can be assessed without regard to text generation.

In this paper, we focus on the semantic representation of the content of an explanation. The ambition is to offer a tool allowing to seamlessly generate textual explanations with NLG techniques in the target language. The challenge is to obtain an abstract semantic representation, i.e. a structure that connects explicitly concepts to each other. This requirement was put forward with natural language generation for decision support (Reiter, 2006). To our knowledge, no such representation has been introduced specifically for explanations.

As the representation will be an input for the text generation and the evaluation processes, it needs to be a coherent structure constructed in a manner that preserves expressiveness and simplicity for being used by XAI applications. Indeed, this structure will play a key role regarding the understanding of the text produced. The literature in cognitive science shows that text production and its understanding are greatly connected (Bos et al., 2015). On the other hand, different aspects should be taken into account while producing an explanation in order to increase user acceptance. For instance, it should be simple, contrastive, adapted to the context, etc. (Miller, 2019). Therefore, the representation needs to consider these elements to be useful.

In addition, a specific task towards the generation of an explanation, is to determine the nature of the pieces of information to involve in an explanation. They are connected to each other by precise relations (e.g. causality) which need to be carefully defined. This subject has been notably studied by cognitive science researchers. They have developed text representation and comprehension models (Kintsch and Van Dijk, 1978; Van den Broek et al., 1999) with a strong focus on narrative representation and comprehension in the 80-90's (Zwaan and Radvansky, 1998). Indeed, narrative text have properties actively sought in cognitive science such as foregrounding the way inferences are generated during reading (Graesser et al., 1991). Some of these models are dedicated to the representation of structured stories, and model situations involving multiple sources of knowledge (e.g. causality, agentivity) with a great expressiveness. The next section is dedicated to discuss some knowledge representations and especially the narrative representation.

3 Background

Historically, the knowledge representation of an explanation was a question tackled during the emergence of expert systems in the 80-90's. The knowledge involved in an explanation was separated into a reasoning knowledge base and a domain knowledge base (Swartout, 1983), and later, the use of a knowledge base dedicated to communication has been also considered (Barzilay et al., 1998). Most of these explanations were represented with conceptual graphs, which are logic-based semantic networks (Sowa, 2000). Indeed, they have demonstrated good properties to represent content with a convenient expressiveness. Most of the models we will now introduce derive from them.

To our knowledge, modern intelligent systems have not defined a way to represent specifically an explanation in a form that highlights the relationships of its constituents. The representation of an explanation must be able to deal with the multi-

ple nature of involved components (e.g. objects, assertions, properties) and relations between them (e.g. causality, spatial or temporal). At the moment, state-of-the-art approaches (Forrest et al., 2018; Alonso and Bugarın, 2019; Pierrard et al., 2019; Baaj and Poli, 2019) use mostly surface realizers like SimpleNLG (Gatt and Reiter, 2009) to produce textual explanations.

There are several drawbacks to use directly a surface realizer. On the one hand, intelligent systems justify their decisions by selecting clues of their reasoning but neither these algorithms nor the realizers take the structure of the textual explanation into account. On the other hand, surface realizers like SimpleNLG use both linguistically and syntactically oriented knowledge representations only to represent the roles of the concepts in the text.

To fill this lack of such representations, we investigated knowledge representations in Natural Language Processing domain which are numerous and evolving from lexically-based to compositionally-based (Cambria and White, 2014). Due to space limitation, we limit our discussion to three approaches, by highlighting the major difficulties with them.

Firstly, we can mention a popular representation, named *conceptual graphs* which are used as schemes for semantic representation of text (e.g. Abstract Meaning Representation (AMR)) (Abend and Rappoport, 2017). Nevertheless, these models are tied to semantic parsing of sentences. For a sentence, approaches like AMR (Banarescu et al., 2013) create a rooted directed acyclic graph, whose relations link the root node to some segments of the sentence. Relationships annotate the role of each segment at the sentence level (Abend and Rappoport, 2017). For instance, to specify a semantic AMR annotates segments of text with specific tags, for instance ":location" or ":time" relations. However, it is not possible to describe with relations higher-level semantic such as an event occurring before another one.

Secondly, many NLP applications use text organization theories such as *Rhetorical Structure Theory (RST)* (Mann and Thompson, 1987) that emphasizes text organization. It consists in aggregating small units of text (Elementary Discourse Units) by linking them with discourse relations (e.g. restatement, purpose). This approach lacks of granularity since it cannot manipulate abstract concepts and their own relations (e.g. subsumption or mereology).

Finally, *ontologies* bring the good level of abstraction and are also used in some NLG systems (Galanis and Androutsopoulos, 2007). However, semantic triples used with modern ontology languages such as OWL are not suitable to express causality or other logical operations which are key elements in explanation (Miller, 2019) (e.g. proposition such as "A and B cause C").

The former three approaches are difficult to deflect from their first purpose. It leads us to explore how text is represented in fields related to NLP. Furthermore, we notice that researchers have recently proposed NLG approaches based on comprehension theories to build a comprehension-driven NLG planner (Thomson et al., 2018). We support and investigate these works, emphasizing that the production of text by AI systems with a focus on comprehension is a promising direction. The next section focuses on narrative representations that are a specific kind of conceptual graphs.

4 Narrative representation and conceptual graph structures

Narrative representation is both studied in AI and cognitive science and consists in modeling the essence of a story that is independent of the audience, the narrator and the context (Elson, 2012b). The literature is abundant and it is difficult to be exhaustive while enumerating narrative representations and their applications, and this is not our aim in this paper.

Among these models, we can distinguish psychology contributions, e.g. Mandler and Johnson's story grammar (Mandler and Johnson, 1977) and Trabasso's causal network (Trabasso and Van Den Broek, 1985), and AI contributions, e.g. conceptual graph structures (Graesser et al., 1991), plot units (Lehnert, 1981), and more recently Story Intention Graphs (Elson, 2012b).

Those different approaches were successfully applied to story variation in NLG (Rishes et al., 2013; Lukin and Walker, 2019), story analogy detection (Elson, 2012a) and question-answering (Graesser and Franklin, 1990; Graesser et al., 1992).

The conceptual graph structures of QUEST (Graesser et al., 1992) have then been extended and applied to new applications such as capturing expert knowledge in biology (Gordon, 1996), or

text representation (Graesser et al., 2001).

Conceptual graph structures are semantic networks in which it is possible to define abstract concepts and formulate statements which makes possible to form causal networks with basic logical inference representation (with "and", "xor", "implies", "causes" and "enables" relations), goal hierarchies, taxonomic hierarchies, spatial structures, and time indexes within a unique framework.

In such graphs, (Graesser et al., 2001) consider five types of nodes:

- concepts (C) are nouns,
- states (S) are unchangeable facts within the time-frame,
- events (E) are episodic propositions,
- goals (G) are statements that an agent wants to achieve, and
- styles (Sy) describe the qualitative manner or intensity of statements.

The semantic network is formed by connecting nodes with the help of a catalogue of twenty-two relations for text representation. Each relation has a definition and a composition rule, and may have synonyms, inverses, sub-types and negation relations. As example, it can represent that the goal "the cat wants to eat" is initiated by the statement "the cat is hungry". Indeed, the relation "initiates" is defined as the initiation of a goal, and is a directed arc from a node that is either a state (S), an event (E) or a style (Sy), to a goal (G) node. It has "elicits" as synonym, "condition", "circumstance" and "situation" are its inverse, and "disables" is its negation. In the next section, we discuss why conceptual graph structures seem to be good candidates for a general explanation representation in XAI.

5 Discussion

We aim at a unified representation of the content of explanations which is independent from the AI model that generates them. Our review of the state-of-the-art revealed the conceptual graph structures for text representation (Graesser et al., 2001) as a good candidate. Indeed, this model can represent complex arrangement of concepts like hierarchies and taxonomies.

Moreover, the situation of an explanation can be expressed spatially and temporally, incorporating definition of concepts that can contain notably agentivity properties (e.g. goals), attributes (e.g. is-a) and that can emphasize contrastive aspects (e.g. opposite, is-not-a, contradicts..).

From this representation, the core-meaning of causality in explanations can be expressed with *enables* and *causes* relations, which underlie deductive, inductive and abducting reasoning in explanations as argued by (Khemlani et al., 2014). Additionally, it also supports propositional calculus operators and thus allows to represent basic logical inference for logic based XAI. In this conceptual graph, relations are also constrained regarding the kind of nodes they can be applied on: this is a great feature to ensure a correct semantic.

Finally, to handle complex explanations, this model offers a support for the representation of the five dimensions of a "mental representation" of a text. Mental representations are a result of cognitive science applied to the text comprehension process, named the *situation model* (Van Dijk et al., 1983). It describes at least five dimensions in memory: time, space, causation, intentionality and protagonist (Zwaan and Radvansky, 1998) that are all representable in conceptual graph structures.

Despite the expressiveness and the conciseness of this model, some relations are still missing like the representation of disjunctions, and the temporal and spatial aspects are still limited compared to existing XAIs. Nevertheless, conceptual graph structures will be a source of inspiration for our future work.

6 Conclusion

In this paper, some benefits of the use of a semantic representation of explanation were introduced. It can help to link research efforts made by XAI researchers, who extract explanations from AI instantiated models and seek to produce textual explanations. As of today, to our knowledge, XAI systems that produce explanations in natural language use in general lexically and syntactically oriented knowledge representations. In this paper, we argued why these formats are not suitable to represent the justifications provided by modern intelligent systems. We investigated text comprehension studies in cognitive science which led to give support for an expressive and simple semantic network used for text representation (Graesser et al., 2001). We believe that this structure can be a basis for a representation of explanation in AI, which could lead to a potential unification of XAI research works.

References

Omri Abend and Ari Rappoport. 2017. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89.

Jose M Alonso and A Bugarın. 2019. Expliclas: Automatic generation of explanations in natural language for weka classifiers. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 660–665. IEEE.

Jose M Alonso, Alejandro Ramos-Soto, Ehud Reiter, and Kees van Deemter. 2017. An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE.

Ismaïl Baaj and Jean-Philippe Poli. 2019. Natural language generation of explanations of fuzzy inference decisions. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 563–568. IEEE.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Regina Barzilay, Daryl McCullough, Owen Rambow, Jonathan DeCristofaro, Tanya Korelsky, and Benoit Lavoie. 1998. A new approach to expert system explanations. Technical report, COGENTEX INC ITHACA NY.

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, page 1.

Lisanne T Bos, Björn B de Koning, Floryt van Wesel, A Marije Boonstra, and Menno van der Schoot. 2015. What can measures of text comprehension tell us about creative text production? *Reading and writing*, 28(6):829–849.

Paul Van den Broek, Michael Young, Yuhtsuen Tzeng, Tracy Linderholm, et al. 1999. The landscape model of reading: Inferences and the online construction of a memory representation. *The construction of mental representations during reading*, pages 71–98.

Erik Cambria and Bebo White. 2014. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 175–182. ACM.

David K Elson. 2012a. Detecting story analogies from annotations of time, action and agency. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative, Istanbul, Turkey*, pages 91–99.

David K Elson. 2012b. *Modeling narrative discourse*. Ph.D. thesis, Columbia University.

European Council. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. *Official Journal of the European Union (OJ)*, 59(1-88):294.

James Forrest, Somayajulu Sripada, Wei Pang, and George Coghill. 2018. Towards making nlg a voice for interpretable machine learning. In *Proceedings of The 11th International Natural Language Generation Conference*, pages 177–182. Association for Computational Linguistics (ACL).

Dimitrios Galanis and Ion Androutsopoulos. 2007. Generating multilingual descriptions from linguistically annotated owl ontologies: the naturalowl system. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 143–146. Association for Computational Linguistics.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Sallie E Gordon. 1996. Eliciting and representing biology knowledge with conceptual graph structures. In *Knowledge Acquisition, Organization, and Use in Biology*, pages 206–225. Springer.

Arthur Graesser, Jonathan M Golding, and Debra L Long. 1991. Narrative representation and comprehension. *Handbook of reading research*, 2:171–205.

Arthur C Graesser and Stanley P Franklin. 1990. Quest: A cognitive model of question answering. *Discourse processes*, 13(3):279–303.

Arthur C Graesser, Sallie E Gordon, and Lawrence E Brainerd. 1992. Quest: A model of question answering. *Computers & Mathematics with Applications*, 23(6-9):733–745.

Arthur C Graesser, Peter Wiemer-Hastings, and Katja Wiemer-Hastings. 2001. Constructing inferences and relations during text comprehension. *Text representation: Linguistic and psycholinguistic aspects*, 8:249–271.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Sangeet S Khemlani, Aron K Barbey, and Philip N Johnson-Laird. 2014. Causal reasoning with mental models. *Frontiers in human neuroscience*, 8:849.

Walter Kintsch and Teun A Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Wendy G Lehnert. 1981. Plot units and narrative summarization. *Cognitive science*, 5(4):293–331.

Stephanie M Lukin and Marilyn A Walker. 2019. A narrative sentence planner and structurer for domain independent, parameterizable storytelling. *Dialogue & Discourse*, 10(1):34–86.

Jean M Mandler and Nancy S Johnson. 1977. Remembrance of things parsed: Story structure and recall. *Cognitive psychology*, 9(1):111–151.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute.

Corrado Mencar and José M Alonso. 2018. Paving the way to explainable artificial intelligence with fuzzy modeling. In *International Workshop on Fuzzy Logic and Applications*, pages 215–227. Springer.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38.

Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*.

James A Overton. 2012. *Explanation in Science*. Ph.D. thesis, The University of Western Ontario.

Régis Pierrard, Jean-Philippe Poli, and Céline Hudelot. 2019. A new approach for explainable multiple organ annotation with few data. In *Proceedings of the Workshop on Explainable Artificial Intelligence (XAI) 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, XAI@IJCAI 2019*, pages 107–113. IJCAI.

Ehud Reiter. 2006. Natural language generation for decision support. Technical report, Department of Computing Science, University of Aberdeen, UK.

Elena Rishes, Stephanie M Lukin, David K Elson, and Marilyn A Walker. 2013. Generating different story tellings from semantic representations of narrative. In *International Conference on Interactive Digital Storytelling*, pages 192–204. Springer.

John F. Sowa. 2000. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks/Cole Publishing Co., Pacific Grove, CA, USA.

William R Swartout. 1983. Xplain: A system for creating and explaining expert consulting programs. *Artificial intelligence*, 21(3):285–325.

Craig Thomson, Ehud Reiter, and Somayajulu Sripada. 2018. Comprehension driven document planning in natural language generation systems. In *Proceedings of The 11th International Natural Language Generation Conference*, pages 371–380. Association for Computational Linguistics (ACL).

Tom Trabasso and Paul Van Den Broek. 1985. Causal thinking and the representation of narrative events. *Journal of memory and language*, 24(5):612–630.

US Council. 2018. Statement on algorithmic transparency and accountability.

Teun Adrianus Van Dijk, Walter Kintsch, and Teun Adrianus Van Dijk. 1983. *Strategies of discourse comprehension*. Academic Press New York.

David Wulf and Valentin Bertsch. 2017. A natural language generation approach to support understanding and traceability of multi-dimensional preferential sensitivity analysis in multi-criteria decision making. *Expert Systems with Applications*, 83:131 – 144.

Rolf A Zwaan and Gabriel A Radvansky. 1998. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162.