Explaining Bayesian Networks in Natural Language: State of the Art and Challenges

Conor Hennessy, Alberto Bugarín
Centro Singular de Investigación
en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{conor.hennesy,alberto.bugarin.diz}@usc.es

Ehud Reiter
University of Aberdeen
e.reiter@abdn.ac.uk

Abstract

In order to increase trust in the usage of Bayesian networks and to cement their role as a model which can aid in critical decision making, the challenge of explainability must be faced. Previous attempts at explaining Bayesian networks have largely focused on graphical or visual aids. In this paper we aim to highlight the importance of a natural language approach to explanation and to discuss some of the previous and state of the art attempts of the textual explanation of Bayesian Networks. We outline several challenges that remain to be addressed in the generation and validation of natural language explanations of Bayesian Networks. This can serve as a research agenda for future work on natural language explanations of Bayesian Networks.

1 Introduction

Despite an increase in the usage of AI models in various domains, the reasoning behind the decisions of complex models may remain unclear to the end user. The inability to explain the reasoning taking of a model is a potential roadblock to their future usage (Hagras, 2018). The model we discuss in this paper is the Bayesian Network (BN). A natural example of the need for explainability can be drawn from the use of diagnostic BNs in the medical field. Accuracy is, of course, highly important but explainability too would be crucial; the medical or other professional, for instance, should feel confident in the reasoning of the model and that the diagnosis provided is reliable, logical, comprehensible and consistent with the established knowledge in the domain and/or his/her experience or intuition. To achieve this level of trust, the inner workings of the BNs must be explained. Take for example the BN presented in Kyrimi et al. (2020) for predicting the likelihood for coagulopathy in patients. To explain a prediction about coagulopathy based on some observed evidences, not only is the most significant evidence highlighted, but also how this evidence affects the probability of coagulopathy through unobserved variables.

While a very useful tool to aid in reasoning or decision making, BNs can be difficult to interpret or counter-intuitive in their raw form. Unlike decision support methods such as decision trees and other discriminative models, we can reason in different directions and with different configurations of variable interactions. Probabilistic priors and the interdependencies between variables are taken into account in the construction (or learning) of the network, making BNs more suited to encapsulate a complex decision-making process (Janssens et al., 2004). On the other hand, this linkage between variables can lead to complex and indirect relationships which impede interpretability. The chains of reasoning can be very long between nodes in the BN, leading to a lack of clarity about what information should be included in an explanation. With an automatic Natural Language Generation (NLG) approach to explaining the knowledge represented and reasoning process followed in a BN, they can be more widely and correctly utilized. We will outline what information can be extracted from a BN and how this has been used to provide explanations in the past. It will be shown how this can be considered a question of content determination as part of an NLG pipeline, such as that discussed by Reiter and Dale (2000), and highlight the state of the art in natural language explanation of BNs. This is the first such review, to the best of our knowledge, that focuses on explaining BNs in natural language.

2 Bayesian Networks

2.1 Overview

Bayesian Networks are Directed Acyclic Graphs where the variables in the system are represented as nodes and the edges in the graph represent the probabilistic relationships between these variables

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 28–33,
Dublin, Ireland, 18 December 2020. ©2020 Association for Computational Linguistics

(Pearl, 1988). Each node in the network has an associated probability table, which demonstrates the strength of the influence of other connected variables on the probability distribution of a node. The graphical component of a BN can be misleading; It may appear counter-intuitive that the information of observing evidence in the child nodes can travel in the opposite direction of directed arrows from parents to children. The direction of the arrows in the graph are intended to demonstrate direction of hypothetical causation; as such, there would be no arrow from symptom to disease. Depending on the structure of the chains connecting variables in the network, dependencies can be introduced or removed, following the rules of d-separation (Pearl, 1988). These rules describe how observing certain evidence may cause variables to become either dependent or independent, a mechanism which may not be obvious or even intuitive for an end user. Describing this concept of dynamically changing dependencies between variables to a user is one of the unique challenges for the explanation of BNs in particular.

It is not only the graphical component of the BNs which can invite misinterpretation; Bayesian reasoning in particular can often be unintuitive; the conditional probability tables themselves may not be interpretable for an average user. Take the example from Eddy (1982) from the medical domain where respondents involved in their study struggled to compute the correct answers to questions where Bayesian reasoning and conditional probability were involved. Examples are given by Keppens (2019); de Zoete et al. (2019) of the use of BNs to correct cases of logical fallacy or to solve paradoxes in the legal field. As these models can provide seemingly counter-intuitive answers, the provision of a convincing mechanism of explanation is crucial.

2.2 What can be Explained?

There are several approaches to extracting and explaining information contained in BNs; A taxonomy was first laid out by Lacave and Díez (2002) for the types of explanations that can be generated. Explanations are said to fall into 3 categories.[1]

- Explanation of the evidence typically amounts to providing the most probable explanation of a node of interest in the network by selecting the configurations of variables that are most likely to have resulted in the available evidence. In BNs this is often done by calculating the maximum a-posteriori probability for the evidence. This can aid in situations such as medical diagnoses and legal cases.
- Explanation of the model involves describing the structure of the network and the relationships contained within it. Unlike other discriminative models such as decision trees, prior probabilities and expert knowledge may have been used to construct the BN and may need to be explained. This can be used to provide domain knowledge for end users or for debugging a model.
- Explanation of the reasoning has the goal of describing the reasoning process in the network which took place to obtain a result. This can also include explanations of why a certain result was not obtained, or counterfactual explanations about results that could be obtained in hypothetical situations (Constantinou et al., 2016).

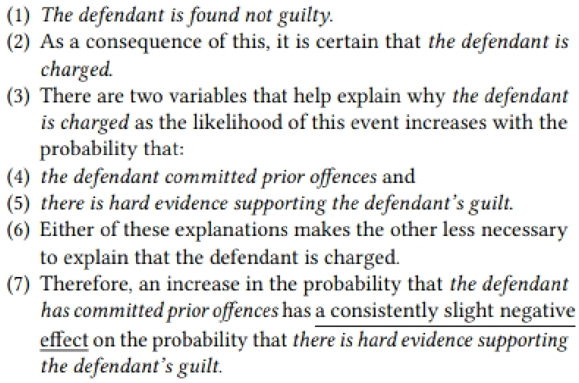
There have been many methodologies suggested to extract content that could be used to generate explanations under all 3 categories (Kyrimi et al., 2020; Lacave et al., 2007). It is crucial to consider the target user when creating explanations of BNs. For example, many previous explanations of BNs to aid in clinical decision support focused on explaining the intricacies of the BN itself, which would be of no interest to a doctor, rather than *using* the information from the BN to offer relevant explanations to aid in medical reasoning. On the other hand, explanations that explicitly describe the model could be useful for developers in the construction of BNs and to aid in debugging when selecting the relevant variables and structure of the model. While the question of what to explain is highly important, so too is how it is explained. This is why the extraction of information from a BN should be viewed as the content determination stage as part of a larger NLG pipeline. In the past, there has been a greater emphasis placed on visual explanations of BNs using graphical aids and visual tools, than with verbal approaches (Lacave and Díez, 2002). This could be due to the unawareness of the benefits of natural language explanations or of the possibility of viewing the extraction of information from a BN as a question of content determination for NLG.

[1] It should be noted that explanation here signifies *what* to explain rather than *how* it should be explained

3 Need for Natural Language Explanation

If generated textual explanations are written for a purpose and an audience, have a narrative structure and explicitly communicate uncertainty, they can be a useful aid in explaining AI systems (Reiter, 2019). In early expert systems, explanation was considered a very important component of the system and textual explanations were identified as a solution for explaining reasoning to users (Shortliffe and Buchanan, 1984).

Textual explanation was also identified as important for the explanation of Bayesian reasoning; Haddawy et al. (1997) claimed that textual explanation would not require the user to know anything about BNs in order to interact with it effectively. Many of the early textual explanations took the form of basic canned text and offered very stiff output. The developers of the early explanation tools for BNs expressed a definite desire for a more natural language approach, rather than outputting numerical, probabilistic information, as well as facilities for interaction and dialog between user and system (Lacave et al., 2007). The state of the art at the time did not allow for the creation of such capabilities for the system, and these challenges have still not been sufficiently revisited with the capability of the state of the art of today.



(1) *The defendant is found not guilty.*
(2) As a consequence of this, it is certain that *the defendant is charged.*
(3) There are two variables that help explain why *the defendant is charged* as the likelihood of this event increases with the probability that:
(4) *the defendant committed prior offences* and
(5) *there is hard evidence supporting the defendant's guilt.*
(6) Either of these explanations makes the other less necessary to explain that the defendant is charged.
(7) Therefore, an increase in the probability that *the defendant has committed prior offences* has a consistently slight negative effect on the probability that *there is hard evidence supporting the defendant's guilt.*

Figure 1: Example of explanation in legal domain from (Keppens, 2019)

Figure 1 contains an example of a potential natural language explanation that could be generated from a BN following the methodology in (Keppens, 2019). This explanation attempts to pacify feelings of guilt in jurors. In the given example, members of a jury may feel regret after, having returned a verdict of not guilty, learning that the accused had prior convictions. By fixing ”non-guilty verdict” and ”prior convictions” as true in the network, the explanation aims to convince a juror that a defendant having prior convictions does not increase the probability of the existence of hard evidence supporting their guilt. While the clarity may suffer due to the explanation in present tense of events that have taken place in different timelines, this example is a marked improvement on past textual explanations of a BN. A narrative is created around the defendant and vague, natural language is used to create arguments to persuade the juror; much more convincing than the common approach of printing observations and probabilistic values.

4 Textual Explanations of BNs

4.1 State of the Art

Several of the earliest attempts of the explanation of BNs were highlighted by Lacave and Díez (2002).This includes early attempts to express Bayesian reasoning linguistically and several systems with rudimentary textual explanations of the model or its reasoning, such as BANTER, B2, DIAVAL and Elvira (Haddawy et al., 1994; Mcroy et al., 1996; Díez et al., 1997; Lacave et al., 2007). In many cases, the state of the art at the time was deemed insufficient to provide satisfactory natural language explanation facilities (Lacave et al., 2007)

More recently, the explanation tool for BNs developed by van Leersum (2015) featured a textual explanation component. While opting for a linguistic explanation of probabilistic relationships and providing a list of arguments for the result of a variable of interest, the language of the templates used to create is more purely a description of the BN rather than providing natural language answers to the problem by using the BN. Such a style of explanation would require a user to have a high level of domain knowledge and even knowledge of how BNs operate. In the legal domain, an approach has been suggested to combine BNs and scenarios which, if combined with NLG techniques, could be used to create narratives to aid in decision making for judge or jury (Vlek et al., 2016).A framework is proposed by Pereira-Fariña and Bugarín (2019) for the explanation of predictive inference in BNs in natural language.

Keppens (2019) also described an approach to the determination of content from a BN as part of the NLG pipeline, using the support graph method described by Timmer et al. (2017). It is then shown how this content is trimmed and ordered at the high-level planning stage. In order to implement the high level-plan, sentence structures are generated at the micro-planning stage.

BARD is a system created to support the collaborative construction and validation of BNs (Nicholson et al., 2020; Korb et al., 2020). As part of this system, a tool for generating textual explanations of relevant BN features was developed, with the view that as BNs become highly complex, they should be able to verbally explain themselves. The tool implements “mix of traditional and novel NLG techniques” and uses common idioms and verbal descriptions for expressing probabilistic relationships. The explanation describes probabilities of target variables if no evidence is entered. When evidence is entered, additional statements are generated about the evidence for the given scenario, and how the probabilities in the model have changed as a result. There is also an option to request a more detailed explanation also containing the structure of the model, how the target probabilities are related to each other, the reliability and bias of the evidence sources, why the evidence sources are structurally relevant and the impact of the evidence items on each hypothesis. The team aims to improve and test the verbal explanations and to add visual aids in the future. The system shows how natural language explanations can be used in the collaborative construction of BNs and this could be extended to provide for a collaborative debugging facility for an existing BN. The interactive explanation capability could be expanded to allow for natural language question and answering between user and system.

A three level approach to the explanation of a medical BN is suggested by Kyrimi et al. (2020) where, given a target variable in the system, a list of significant evidence variables, the flow of information through intermediate variables between target and evidence and the impact of the evidence variables on intermediate variables are explained. The verbal output uses templates to create textual and numerical information structured in simple bullet points.The small-scale evaluation of the explanation by participating clinicians produced mixed opinions.The explanations were evaluated based on similarity to expert explanations, increase of trust in model, potential clinical benefit and clarity. The team acknowledged several limitations of the study, and while failing to demonstrate an impact on trust, they did show the clarity and similarity of the explanation to clinical reasoning, and that it had an affect on clinician’s assessment.

4.2 Discussion and Challenges for Future Work

There is still much work to be done to achieve automatic generation of natural language explanations of BNs. This includes further examination of what information should be extracted from BNs for explanatory purposes, and how that information should be presented:

- Within the content determination stage, there is still a lack of clarity about what information from the BN is best to communicate to users. Based on the communicative goals of an explanation, and following the taxonomy for explanation introduced by Lacave and Díez (2002), the appropriate content should be extracted. Furthermore, greater consideration should be given to the goals and target of an explanation in the planning stage.

- The literature has focused on the content determination stage of the NLG process. There is less work on the planning stages and less still on realisation, particularly in real use cases or domains.

- It appears that the majority of verbal explanation of BNs are generated by the gap-filling of templates. This rigid approach does not lend itself to the dynamic nature of BNs. Templates are generally written in present tense which can may lead to confusing explanations, as the evidences are often observed in different timelines. The dynamic generation of textual explanation is not commonly considered and we have been unable to find any corpus to train a model for the explanation of BNs. Furthermore, to our knowledge no end-to-end NLG approaches for generating textual descriptions of BN from data have been presented in the literature.

- There are relatively few methods discussing a story or narrative-style approach to explanation. For BNs, this approach seems to only have been considered in the legal domain, despite recognition as an effective means of explanation in general (Reiter, 2019).

- Past work on the linguistic expression of probabilistic values is often not considered. Developers commonly opt to print numerical values leading to less acceptable explanations.

There are several challenges related to enriching the potential for explanation in existing and future BN systems:

- Related work on enriching the ability for causal inference with BNs would allow for causal attributions in explanations, which is clearer for people than the language of probabilistic relationships (Biran and McKeown, 2017).

- The desire expressed in the past for the capability of a user-system natural language dialogue facility has also not been addressed (Lacave et al., 2007). This could be used as an education tool for students, as suggested by Mcroy et al. (1996). Users in non-technical domains such as medicine and law may wish to interact with Bayesian systems in the same way they would with experts in their respective domains, getting comprehensible insights about the evidences that support the conclusions produced by a Bayesian model.

- Natural language explanation methods could be integrated with BN-based systems and tools currently being applied successfully in industry, such as those in healthcare technology companies, to aid developers and increase their value for end users (McLachlan et al.).

Finally, there is related work remaining in order to sufficiently evaluate the output of any explanation facility for a BN:

- Many of the explanations that have been generated have not been comprehensively validated to be informative or useful. Intrinsic and extrinsic evaluations should be conducted both by humans and using state of the art automatic metrics where appropriate. Determining how best to evaluate textual explanations of a BN will be a crucial component for their more widespread use in the future (Barros, 2019; Reiter, 2018).

- It should be evaluated how natural language explanations compare with visual explanations and in which situations a particular style (or a combination of both) should be favoured.

5 Conclusion

It is clear that in the 1990's and early 2000's, there was a desire for implementing an effective natural language explanation facility for BNs. In many cases, the previous attempts were deemed unsatisfactory by their developers or evaluators, due to the fact that the state of the art at the time limited their ability to provide the kind of natural explanations that they wished. This paper highlights several challenges which should be revisited with state of the art NLG capabilities and with the improved ideas we now have of what should be provided in a satisfactory explanation.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 860621. It was also funded by the Spanish Ministry for Science, Innovation and Universities, the Galician Ministry of Education, University and Professional Training and the European Regional Development Fund (grants TIN2017-84796-C2-1-R, ED431C2018/29 and ED431G2019/04).

References

C. Barros. 2019. *Proposal of a Hybrid Approach for Natural Language Generation and its Application to Human Language Technologies*. Ph.D. thesis, Department of Software and Computing systems, Universitat d'Alacant.

Or Biran and Kathleen McKeown. 2017. Human-Centric Justification of Machine Learning Predictions. In *Proceedings of the Twenty-Sixth International Journal Joint Conferences on Artificial Intelligence*, IJCAI 2017, pages 1461–1467.

Anthony Costa Constantinou, Barbaros Yet, Norman Fenton, Martin Neil, and William Marsh. 2016. Value of information analysis for interventional and counterfactual Bayesian networks in forensic medical sciences. *Artificial Intelligence in Medicine*, 66:41–52.

F. J. Díez, J. Mira, E. Iturralde, and S. Zubillaga. 1997. Diaval, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, 10(1):59–73.

David M. Eddy. 1982. Probabilistic reasoning in clinical medicine: Problems and opportunities. In *Judgment under Uncertainty: Heuristics and Biases*, pages 249–267. Cambridge University Press.

P. Haddawy, J. Jacobson, and C. E. Kahn. 1997. BANTER: A Bayesian network tutoring shell. *Artificial Intelligence in Medicine*, 10(2):177–200.

P. Haddawy, J. Jacobson, and C.E. Kahn. 1994. An educational tool for high-level interaction with Bayesian networks. In *Proceedings Sixth International Conference on Tools with Artificial Intelligence. TAI 94*, pages 578–584.

H. Hagras. 2018. Toward human-understandable, explainable AI. *Computer*, 51(9):28–36.

Davy Janssens, Geert Wets, Tom Brijs, Koen Vanhoof, Theo Arentze, and Harry Timmermans. 2004. Improving performance of multiagent rule-based model for activity pattern decisions with bayesian networks. *Transportation Research Record*, 1894(1):75–83.

Jeroen Keppens. 2019. Explainable Bayesian Network Query Results via Natural Language Generation Systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL '19, pages 42–51. Association for Computing Machinery.

Kevin B. Korb, Erik P. Nyberg, Abraham Oshni Alvandi, Shreshth Thakur, Mehmet Ozmen, Yang Li, Ross Pearson, and Ann E. Nicholson. 2020. Individuals vs. BARD: Experimental evaluation of an online system for structured, collaborative bayesian reasoning. *Frontiers in Psychology*, 11:1054.

Evangelia Kyrimi, Somayyeh Mossadegh, Nigel Tai, and William Marsh. 2020. An incremental explanation of inference in Bayesian networks for increasing model trustworthiness and supporting clinical decision making. *Artificial Intelligence in Medicine*, 103:101812.

Carmen Lacave and Francisco J. Díez. 2002. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127.

Carmen Lacave, Manuel Luque, and Francisco Javier Diez. 2007. Explanation of Bayesian Networks and Influence Diagrams in Elvira. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(4):952–965.

J. van Leersum. 2015. Explaining the reasoning of bayesian networks with intermediate nodes and clusters. Master's thesis, Faculty of Science, Universiteit Utrecht.

Scott McLachlan, Kudakwashe Dube, Graham A Hitman, Norman E Fenton, and Evangelia Kyrimi. Bayesian networks in healthcare: Distribution by medical condition. *Artificial Intelligence in Medicine*, 107:101912.

Susan W. Mcroy, Alfredo Liu-perez, James Helwig, and Susan Haller. 1996. B2: A tutoring shell for bayesian networks that supports natural language interaction. In *In Working Notes, 1996 AAAI Spring Symposium on Artificial Intelligence and Medicine*, pages 114–118.

Ann E. Nicholson, Kevin B. Korb, Erik P. Nyberg, Michael Wybrow, Ingrid Zukerman, Steven Mascaro, Shreshth Thakur, Abraham Oshni Alvandi, Jeff Riley, Ross Pearson, Shane Morris, Matthieu Herrmann, A. K. M. Azad, Fergus Bolger, Ulrike Hahn, and David Lagnado. 2020. BARD: A structured technique for group elicitation of Bayesian networks to support analytic reasoning. *arXiv e-prints*, page arXiv:2003.01207.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc.

Martín Pereira-Fariña and Alberto Bugarín. 2019. Content Determination for Natural Language Descriptions of Predictive Bayesian Networks. In *11th Conference of the European Society for Fuzzy Logic and Technology*, EUSFLAT 2019, pages 784–791. Atlantis Press.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter. 2019. Natural Language Generation Challenges for Explainable AI. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, pages 3–7. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Edward H. Shortliffe and Bruce G Buchanan. 1984. *Rule-Based Expert System – The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA.

Sjoerd T. Timmer, John-Jules Ch. Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. 2017. A two-phase method for extracting explanatory arguments from Bayesian networks. *International Journal of Approximate Reasoning*, 80:475–494.

Charlotte S. Vlek, Henry Prakken, Silja Renooij, and Bart Verheij. 2016. A method for explaining Bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law*, 24(3):285–324.

Jacob de Zoete, Norman Fenton, Takao Noguchi, and David Lagnado. 2019. Resolving the so-called "probabilistic paradoxes in legal reasoning" with Bayesian networks. *Science & Justice*, 59(4):367–379.