Towards Harnessing Natural Language Generation to Explain Black-box Models

Ettore Mariotti, Jose M. Alonso
Centro Singular de Investigación en
Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago
de Compostela, Spain
{ettore.mariotti,josemaria.alonso.moral}@usc.es

Albert Gatt
Institute of Linguistics
and Language Technology,
University of Malta (UM)
albert.gatt@um.edu.mt

Abstract

The opaque nature of many machine learning techniques prevents the widespread adoption of powerful information processing tools for high stakes scenarios. The emerging field of Explainable Artificial Intelligence aims at providing justifications for automatic decision-making systems in order to ensure reliability and trustworthiness in users. To achieve this vision, we emphasize the importance of a natural language textual explanation modality as a key component for a future intelligent interactive agent. We outline the challenges of explainability and review a set of publications that work in this direction.

1 Introduction

In recent times the use of Machine Learning (ML) has changed many fields across a wide range of domains, revealing the potential for an information processing revolution in our society (West, 2018). Even though there already exist many commercial applications that use ML for delivering products, these are limited by the often opaque nature of the underlying models (Goodman and Flaxman, 2017).

In fact, to produce highly predictive models that reach high-performance metrics on given tasks, commercial products often end up with models whose behavior and rationale in making decisions are not clearly understandable by humans.

This is a big issue in all those applications where trust and accountability in the prediction have the highest priority like healthcare, military, finance, or autonomous vehicles.

This need for explainable models has made many big institutions, including the European Union (Hamon et al., 2020), and the US Defense Advanced Research Projects Agency (DARPA) (Gunning and Aha, 2019) push for funding research in eXplainable Artificial Intelligence (XAI), a relatively new and very active research area with the aim of providing human insight into the behavior of information-processing tools.

The three main XAI challenges are: (1) designing explainable models; (2) implementing explanation interfaces; and (3) measuring the effectiveness of the generated explanations.

Of the many ways of presenting an explanation, natural language is particularly attractive as it allows people with diverse backgrounds and knowledge to interpret it (Alonso et al., 2020), thus potentially allowing the interested end-user to understand the model without requiring a detailed background in mathematics and information engineering. This is a mandatory step if we want to make these tools available to the non-technical wider population. The goal of this paper is to provide a general overview of tools and approaches for providing linguistic explanations of ML models to general users.

The rest of the paper is organized as follows. In section 2 we present a brief overview of XAI field and its challenges. In section 3 we explore how XAI can integrate with Natural Language Generation (NLG). Finally, we summarize the main conclusions in section 4.

2 Open Challenges in XAI

As mentioned in the introduction, XAI faces three main challenges: models, interfaces and evaluations. In this section, we provide a high-level overview of each of them.

2.1 Designing Explainable Models

Different kinds of models provide different explanations. As a first approximation we can distinguish between classes of models depending on their intrinsic ability to be meaningfully inspected. We can picture this taxonomy with a block diagram as shown in Fig. 1.

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 22–27,
Dublin, Ireland, 18 December 2020. ©2020 Association for Computational Linguistics

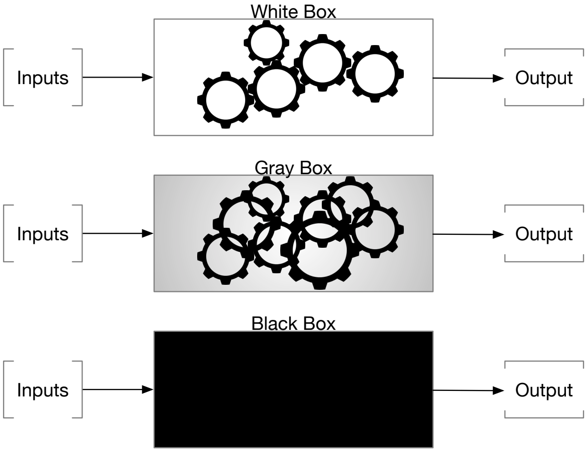

Figure 1: A block diagram representation of different models. White boxes have a clear and decomposable internal representation and processing. Gray boxes are still decomposable but understanding them is less straightforward. For black boxes, processing is assumed unknown and only the input-output behaviour can be inspected.

2.1.1 White-box Models

White models, sometimes called "transparent" models (Barredo Arrieta et al., 2020), are those that behave in a way that humans can understand conceptually and for which their processing can be decomposed to some extent into meaningful and understandable ways. The idea is that we can "see through" them in a block diagram and inspect their functioning. They are easier to be explained but typically reach lower performances on complex tasks. Examples of those include linear models, decision trees, nearest neighbors, rule-based learners and general additive models.

2.1.2 Gray-box Models

Gray boxes are models whose internal structure can be inspected, but for which clear explanations are more difficult to produce. This is because they rely on formalisms, such as probability and plausibility, which differ from perspectives humans find more intuitive (e.g., Bayesian networks or fuzzy systems). These models lack a crisp internal representation which can be displayed in a categorical fashion and instead use soft thresholds and/or conditional probabilities. In this regard Eddy (1982) and Elsaesser (1987) show how people have difficulty with interpreting probabilistic reasoning, especially when it is described numerically.

2.1.3 Black-box Models

With "black box" we refer to those models whose behavior is not directly understandable. Some publications deal with "opening the box", digging into the specific construction details of a class of models by decomposing their whole processing structure into smaller understandable parts and intermediate representations (Olah et al., 2018) or by trying to infer the contribution of each feature to the final outcome (thus effectively "grayifying" them) (Montavon et al., 2018). Others instead "leave the box closed", ignoring the internals of the model, and restrict their scrutiny to the relationships between inputs and outputs. The literature refers to the latter approach as "post-hoc", meaning that the explanation process is decoupled from the inference process, and might not actually represent the real computation happening, but is rather a human readable justification of what is happening (Lipton, 2018). Some examples of black boxes are tree ensembles (e.g., random forests), support vector machines, multi-layer neural networks, convolutional neural networks, recurrent neural networks and generative adversarial networks.

2.2 Implementing Explanation Interfaces

Given a model and a prediction the next problem is to provide an interface that is able to produce a meaningful explanation. The issue is to try to understand what is the best explanation to provide to the user. "What is an explanation?" is a question that has puzzled philosophers, psychologists, and social scientists long before the engineering community stepped into the scene. A great heritage that we can distill from these previous works is that explanations are narratives of causal relationships between events. But it is also clear that while a certain event may have a very long causal history, an explainee (i.e., one who receives the explanation) might consider relevant only a small subset of this history, depending on his/her personal cognitive biases (Miller, 2019). This highlights the fact that different people might judge more relevant different explanations given their different interests or background. Thus a good explanation is dependent on who is going to receive it. But this also points to the fact that explanation is a process, a dialogue between explainer and explainee, rather than a one-shot result.

Various XAI methods have been developed to answer specific one-shot questions, including:

- "Why was this class predicted instead of that?": counterfactual (Russell, 2019),
- "How did each feature contribute to this prediction?": feature importance (Lundberg and Lee, 2017; Fisher et al., 2019)
- "Which data points in your training contributed to the outcome?": explanation by example (Kanehira and Harada, 2019)
- "What happens if I slightly change this input?": local explanation (Goldstein et al., 2015)
- "What is the minimal change in the input required to produce this particular result?": counterfactual and local (Guidotti et al., 2018)

Unfortunately, as far as we know, little to no attention was given so far to an interactive system that could adapt to the user needs and provide "the most effective" explanation for a given situation.

We suggest that a natural language interface between the user and an explanation agent (also supported by visualization techniques) will be a necessary key step toward the trustworthiness and explainability of decision-making systems for high stakes scenarios.

We can imagine a dialogue between a user (**U**) who applied for a loan and an **AI** that rejected it:
U: "Why did I get rejected?"
AI: "Our model predicted that you would be likely to default with a probability of 80%"
U: "Where does that probability come from?"
AI: "For an average user the probability of default is 60%, but the fact that you have less than $50000 and that you are unemployed increase the risk significantly"
U: "What should I do to be granted the loan?"
AI: "If you would got a job and open another account your probability of default would lower to 30% and you would be granted the loan"

2.3 Evaluating Explanation Systems

There is an ongoing discussion in the XAI community on how to evaluate explanation systems. Human assessment is deemed the most relevant, and care should be given in measuring the goodness of an explanation in terms of whether the user understands the model better after the explanation was given (Hoffman et al., 2018). The work of Mohseni et al. (2020) proposes a layered evaluation framework, where the ML algorithm, the explaining interface and global system goals can be better refined for the particular problem at hand and for which specific metric should be constructed.

On the other hand, Herman (2017) points out that excessive reliance on human evaluation could bias the system to be more persuasive rather than transparent due to the user preference of simplified explanations. Quantitative automatic metrics have been for example proposed for evaluating saliency maps for image (Montavon et al., 2017) and text (Arras et al., 2017) classifiers. As will be discussed in section 3.2, Park et al. (2018) propose a dataset labeled with humanly annotated explanations and attentions maps.

All in all, further work is needed for standardizing a general evaluation procedure.

3 Explaining with Natural Language

An explanation can be laid out using different modalities. The general trend in the literature is to represent results in a graphical visual form, but some researchers are using natural language and measuring an increased benefit for the end-user. NLG-based approaches fall into two broad categories: template-based and end-to-end generation.

3.1 Template-based Generation

By leveraging knowledge about the kind of explanation produced about the system it is possible to structure templates that present the output in textual form. The popular LIME method (Ribeiro et al., 2016), which provides a linear approximation of the feature contribution to the output, can be presented in natural language using paragraphs (Forrest et al., 2018), for example with the SimpleNLG toolbox (Gatt and Reiter, 2009). ExpliClas (Alonso and Bugarin, 2019) is a web-service that provides local and global explanations for black boxes by leveraging post-hoc techniques (such as gray model surrogates) in natural language using the NLG pipeline proposed by Reiter and Dale (2000). In the medical domain, a fracture-detecting model has been extended to produce a textual explanation that follows a limited vocabulary and a fixed sentence length (Gale et al., 2018). The authors measured a significant increase in the trustworthiness from a medical population for the textual modality over the visual. While output with templates is easier to control, its static nature sometimes produces sentences that are non-natural and lack variation.

3.2 End-to-end Generation

With the use of a large corpus of humanly labeled data-to-text it is possible to generate sentences without specifying a template a priori. The computer-vision community leveraged the machine translation encoder-decoder framework in order to create systems that are able to semantically describe where and what was detected by an image-classification model (Xu et al., 2015). In Zhang et al. (2019) an image caption model was trained on image-pathologist report pairs in order to produce an automatic textual report as an intermediate step for an interpretable whole-slide cancer diagnosis system. In Hendricks et al. (2016) a model is trained with both an image and a textual description of its content in order to produce an object prediction and a textual justification. The introduction of visual question-answering (VQA-X) and activity recognition (ACT-X) labeled with humanly annotated textual justification and visual segmentation of the relevant parts of the image (Park et al., 2018) allowed to train models that jointly explain a prediction with both text and a visual indication of the relevant portion of the input. This approach is on the other hand expensive (data collection and model training) and occasionally might provide incoherent explanation while being vulnerable to adversarial attacks (Camburu et al., 2020).

3.3 Evaluating Natural Language Generation

The work of van der Lee et al. (2019) highlights an open debate in the NLG community for finding the right way to measure the goodness of generated texts. The main issues revolve around the following questions:

1. Is it possible to rely on automatic metrics only?
2. How should human evaluation be done?

Moreover, there is a significant divergence in how different papers define concepts like “fluency” and “adequacy”.

Textual explanations should first of all be readable (well written, natural, consistent, etc.), but they also need to be effective and useful for the end-user. While automatic metrics such as BLEU, METEOR and ROUGE are quick, repeatable and cheap techniques for roughly assessing language quality, Belz and Reiter (2006), Reiter and Belz (2009) and Reiter (2018) point out that these metrics might not adequately measure quality of content. In addition, Post (2018) shows how different libraries have different default values for the parameters used in computing automatic metrics, thus making comparisons across different publications more difficult. More importantly, automatic metrics have been observed to not correlate with human evaluations (Novikova et al., 2017). That said, while human evaluation remains the gold standard for the general assessment of overall system quality, using it at every step of the development process would be too expensive and slow (van der Lee et al., 2019).

So, goodness of text generated is a prerequisite but is not enough in the context of XAI. New evaluation protocols and best practices in NLG for XAI need to be defined and agreed upon by the scientific community, as this will enable fair comparisons between systems and foster technological improvement.

4 Conclusions

XAI is an emerging field that aims to providing explanations for decision tools that will enable them to gain trust in their users and their wide adoption by the market. In order to achieve this, textual explanations are essential but to date few works have directly addressed this possibility.

Current trends in explainability push toward making intrinsically more interpretable models or in making opaque models more understandable. There is no agreed-upon definition of explanation and further theoretical work should try to bridge the gap between the large corpus of theoretical speculation coming from social sciences and the empirical work pursued in Artificial Intelligence.

This as yet ill-defined nature of the task leaves much work to do in the standardization of processes for measurement of explanation effectiveness. In this regard, both objective and subjective measures should be considered, especially if evaluation involves human participants.

Moreover, since the explanation process is dependent on who is receiving the explanation, we envision an interactive agent that is able to dialogue with the user. From this perspective, the NLG community can contribute significantly to this goal by providing a linguistic layer to the many XAI methods being proposed so far.

Acknowledgments

This research is carried out in the framework of the NL4XAI project which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. Jose M. Alonso is *Ramón y Cajal* Researcher (RYC-2016-19802). His research is supported by the Spanish Ministry of Science, Innovation and Universities (grants RTI2018-099646-B-I00, TIN2017-84796-C2-1-R, TIN2017-90773-REDT, and RED2018-102641-T) and the Galician Ministry of Education, University and Professional Training (grants ED431F 2018/02, ED431C 2018/29, ED431G/08, and ED431G2019/04). These grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

J.M. Alonso, S. Barro, A. Bugarin, K. van Deemter, C. Gardent, A. Gatt, E. Reiter, C. Sierra, M. Theune, N. Tintarev, H. Yano, and K. Budzynska. 2020. Interactive Natural Language Technology for Explainable Artificial Intelligence. In *1st Workshop on Foundations of Trustworthy AI integrating Learning, Optimisation and Reasoning (TAILOR), at the European Conference on Artificial Intelligence (ECAI)*, Santiago de Compostela, Spain.

J.M. Alonso and A. Bugarin. 2019. ExpliClas: Automatic Generation of Explanations in Natural Language for Weka Classifiers. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek. 2017. What is relevant in a text document?: An interpretable machine learning approach. *PLOS ONE*, 12(8):e0181142. Publisher: Public Library of Science.

A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

A. Belz and E. Reiter. 2006. Comparing Automatic and Human Evaluation of NLG Systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

O.-M. Camburu, B. Shillingford, P. Minervini, T. Lukasiewicz, and P. Blunsom. 2020. Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations. ArXiv:1910.03065.

D.M. Eddy. 1982. Probabilistic reasoning in clinical medicine: Problems and opportunities. In Amos Tversky, Daniel Kahneman, and Paul Slovic, editors, *Judgment under Uncertainty: Heuristics and Biases*, pages 249–267. Cambridge University Press, Cambridge.

C. Elsaesser. 1987. Explanation of probabilistic inference for decision support systems. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 394–403, Arlington, Virginia, USA. AUAI Press.

A. Fisher, C. Rudin, and F. Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

J. Forrest, S. Sripada, W. Pang, and G. Coghill. 2018. Towards making NLG a voice for interpretable Machine Learning. In *Proceedings of the 11th International Conference on Natural Language Generation (INLG)*, pages 177–182, Tilburg University, The Netherlands. Association for Computational Linguistics.

W. Gale, L. Oakden-Rayner, G. Carneiro, A.P. Bradley, and L.J. Palmer. 2018. Producing radiologist-quality reports for interpretable artificial intelligence. ArXiv:1806.00340.

A. Gatt and E. Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, pages 90–93, Athens, Greece. Association for Computational Linguistics.

A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. 2015. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65. Publisher: Taylor & Francis.

B. Goodman and S. Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, 38(3):50–57.

R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. ArXiv:1805.10820.

D. Gunning and D. Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58.

R. Hamon, H. Junklewitz, and I. Sanchez. 2020. *Robustness and explainability of Artificial Intelligence: from technical to policy solutions.* Publications Office, LU.

L.A. Hendricks, R. Hu, T. Darrell, and Z. Akata. 2016. Generating Visual Explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19. Springer.

B. Herman. 2017. The Promise and Peril of Human Evaluation for Model Interpretability. In *Proceedings of the NIPS conference*.

R.R. Hoffman, S.T. Mueller, G. Klein, and J. Litman. 2018. Metrics for Explainable AI: Challenges and Prospects.

A. Kanehira and T. Harada. 2019. Learning to Explain With Complemental Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8595–8603, Long Beach, CA, USA. IEEE.

C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, and E. Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Z.C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, 16(3).

S.M. Lundberg and S.-I. Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

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

S. Mohseni, N. Zarei, and E.D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ArXiv:1811.11839.

G. Montavon, S. Bach, A. Binder, W. Samek, and K.-R. Müller. 2017. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition*, 65:211–222.

G. Montavon, W. Samek, and K.-R. Müller. 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing*, 73:1–15.

J. Novikova, O. Dušek, A. Cercas Curry, and V. Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. 2018. The Building Blocks of Interpretability. *Distill*, 3(3):e10.

D.H. Park, L.A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

M. Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

E. Reiter. 2018. A Structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

E. Reiter and A. Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558. Publisher: MIT Press.

E. Reiter and R. Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, Cambridge.

M.T. Ribeiro, S. Singh, and C. Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, New York, USA. Association for Computing Machinery.

C. Russell. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, Atlanta, USA.

D.M. West. 2018. *The Future of Work: Robots, AI, and Automation*. Brookings Institution Press.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the International Conference on Machine Learning (PMLR)*.

Z. Zhang, P. Chen, M. McGough, F. Xing, C. Wang, M. Bui, Y. Xie, M. Sapkota, L. Cui, J. Dhillon, N. Ahmad, F.K. Khalil, S.I. Dickinson, X. Shi, F. Liu, H. Su, J. Cai, and L. Yang. 2019. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245. Publisher: Nature Publishing Group.