Natural Language Generation Challenges for Explainable AI

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

Abstract

Good quality explanations of artificial intelligence (XAI) reasoning must be written (and evaluated) for an explanatory purpose, targeted towards their readers, have a good narrative and causal structure, and highlight where uncertainty and data quality affect the AI output. I discuss these challenges from a Natural Language Generation (NLG) perspective, and highlight four specific "NLG for XAI" research challenges.

1 Introduction

Explainable AI (XAI) systems (Biran and Cotton, 2017; Gilpin et al., 2018) need to explain AI reasoning to human users. If the explanations are presented using natural languages such as English, then it is important that they be accurate, useful, and easy to comprehend. Ensuring this requires addressing challenges in Natural Language Generation (NLG) (Reiter and Dale, 2000; Gatt and Krahmer, 2018).

Figure 1 gives an example of a human-written explanation of the likelihood of water or gas being close to a proposed oil well; I chose this at random from many similar explanations in a Discovery Evaluation Report (Statoil, 1993) produced for an oil company which was deciding whether to drill a well. Looking at this report, it is clear that

- It is *written for a purpose* (helping the company decide whether to drill a well), and needs to evaluated with this purpose in mind. For example, the presence of a small amount of water would not impact the drilling decision, and hence the explanation is not "wrong" if a small amount of water is present.

- It is *written for an audience*, in this case specialist engineers and geologists, by us-

---

It is also unlikely that a water or gas contact is present very close to the well. During the DST test, the well produced only minor amounts of water. No changes in the water content or in the GOR of the fluid were observed. However, interpretation of the pressure data indicates pressure barriers approximately 65 and 250m away from the well [...] It is therefore a possibility of a gas cap above the oil. On the other hand, the presence of a gas cap seems unlikely due to the fact that the oil itself is undersaturated with respect to gas (bubble point pressure = 273 bar, reservoir pressure = 327.7 bar)

Figure 1: Example of a complex explanation

---

ing specialist terminology which is appropriate for this group, and also by using vague expressions (e.g., "minor amount") whose meaning is understood by this audience. A report written about oil wells for the general public (such as NCBPDeepwaterHorizonSpill (2011)) uses very different phrasing.

- It has a *narrative structure*, where facts are linked with causal, argumentative, or other discourse relations. It is not just a list of observations.

- It explicitly *communicates uncertainty*, using phrases such as "possibility" and "unlikely",

If we want AI reasoning systems to be able to produce good explanations of complex reasoning, then these systems will also need to adapt explanations to be suitable for a specific purpose and user, have a narrative structure, and communicate uncertainty. These are fundamental challenges in NLG.

2 Purpose and Evaluation

A core principle of NLG is that generated texts have a *communicative goal*. That is, they have a purpose such as helping users make decisions (perhaps the most common goal), encouraging users to change their behaviour, or entertaining users. Evaluations of NLG systems are based on how well they achieve these goals, as well as the accuracy and fluency of generated texts. Typically we either directly measure success in achieving the goal or we ask human subjects how effective they think the texts will be at achieving the goal (Gkatzia and Mahamood, 2015).

Real-world explanations of AI systems similarly have purposes, which include

- Helping developers *debug* their AI systems. This is not a common goal in NLG, but seems to be one of the most common goals in Explainable AI. The popular LIME model (Ribeiro et al., 2016), for example, is largely presented as a way of helping ML developers choose between models, and also improve models via feature engineering.

- Helping users detect mistakes in AI reasoning (*scrutability*). This is especially important when the human user has access to additional information which is not available to the AI system, which may contradict the AI recommendation. For example, a medical AI system which only looks at the medical record cannot visually observe the patient; such observations may reveal problems and symptoms which the AI is not aware of.

- Building *trust* in AI recommendations. In medical and engineering contexts, AI systems usually make recommendations to doctors and engineers, and if these professionals accept the recommendations, they are liable (both legally and morally) if anything goes wrong. Hence systems which are not trusted will not be used.

The above list is far from complete, for example Tintarev and Masthoff (2012) also include Transparency, Effectiveness, Persuasiveness, Efficiency, and Satisfaction in their list of possible goals for explanations.

Hence, when we evaluate an explanation system, we need to do so in the context of its purpose.

As with NLG in general, we can evaluate explanations at different levels of rigour. The most popular evaluation strategy in NLG is to show generated texts to human subjects and ask them to rate and comment on the texts in various ways. This is leads to my first challenge

- *Evaluation Challenge*: Can we get reliable estimates of scrutabilty, trust (etc) by simply asking users to read explanations and estimate scrutability (etc)? What experimental design (subjects, questions, etc) gives the best results? Do we need to first check explanations for accuracy before doing the above?

Other challenges include creating good experimental designs for task-based evaluation, such as the study Biran and McKeown (2017) did to assess whether explanations improved financial decision making because of increased scrutability; and also exploring whether automatic metrics such as BLEU (Papineni et al., 2002) give meaningful insights about trust, scrutability, etc.

3 Appropriate Explanation for Audience

A fundamental principle of NLG is that texts are produced for users, and hence should use appropriate content, terminology, etc for the intended audience (Paris, 2015; Walker et al., 2004). For example, the Babytalk systems generated very different summaries from the same data for doctors (Portet et al., 2009), nurses (Hunter et al., 2012), and parents (Mahamood and Reiter, 2011).

Explanations should also present information in appropriate ways for their audience, using features, terminology, and content that make sense to the user (Lacave and Díez, 2002; Biran and McKeown, 2017). For example, a few years ago I helped some colleagues evaluate a system which generated explanations for an AI system which classified leaves (Alonso et al., 2017). We showed these explanations to a domain expert (Professor of Ecology at the University of Aberdeen), and he struggled to understand some explanations because the features used in these explanation were not the ones that he normally used to classify leaves.

Using appropriate terminology (etc) is probably less important if the goal of the explanation is debugging, and the user is the machine learning engineer who created the AI model. In this case, the engineer will probably be very familiar

with the features (etc) used by the model. But if explanations are intended to support end users by increasing scrutability or trust, then they need to be aligned with the way that users communicate and think about the problem.

This relates to a number of NLG problems, and I would like to highlight the below as my second challenge:

- *Vague Language Challenge:* People naturally think in qualitative terms, so explanations will be easier to understand if they use vague terms (Van Deemter, 2012) such as "minor amount" (in Figure 1) when possible. What algorithms and models can we use to guide the usage of vague language in explanations, and in particular to avoid cases where the vague language is interpreted by the user in an unexpected way which decreases his understanding of the situation?

There are of course many other challenges in this space. At the content level, it would really help if we could prioritise messages which are based on features and concepts which are familiar to the user. And at the lexical level, we should try to select terminology and phrasing which make sense to the user.

4 Narrative Structure

People are better at understanding symbolic reasoning presented as a narrative than they are at understanding a list of numbers and probabilities (Kahneman, 2011). "John smokes, so he is at risk of lung cancer" is easier for us to process than "the model says that John has a 6% chance of developing lung cancer within the next six years because he is a white male, has been smoking a pack a day for 50 years, is 67 years old, does not have a family history of lung cancer, is a high school graduate [etc]". But the latter of course is the way most computer algorithms and models work, including the one I used to calculate John's cancer risk[1]. Indeed, Kahneman (2011) points out that doctors have been reluctant to use regression models for diagnosis tasks, even if objectively the models worked well, because the type of reasoning used in these models (holistically integrating evidence from a large number of features) is not one they are cognitively comfortable with.

The above applies to information communicated linguistically. In contexts that do not involve communication, people are in fact very good at some types of reasoning which involve holistically integrating many features, such as face recognition. I can easily recognise my son, even in very noisy visual contexts, but I find it very hard to describe him in words in a way which lets other people identify him.

In any case, linguistic communication is most effective when it is structured as a narrative. That is, not just a list of observations, but rather a selected set of key messages which are linked together by causal, argumentative, or other discourse relations. For example, the most accurate way of explaining a smoking risk prediction based on regression or Bayesian models is to simply list the input data and the models result.

> "John is a white male. John has been smoking a pack a day for 50 years. John is 67 years old. John does not have a family history of lung cancer. John is a high school graduate. John has a 6% chance of developing lung cancer within the next 6 years."

But people will probably understand this explanation better if we add a narrative structure do it, perhaps by identifying elements which increase or decrease risks, and also focusing on a small number of key data elements (Biran and McKeown, 2017).

> "John has been smoking a pack a day for 50 years, so he may develop lung cancer even though he does not have a family history of lung cancer."

This is not the most accurate way of describing how the model works (the model does not care whether each individual data element is "good" or "bad"), but it probably is a better explanation for narrative-loving humans.

In short, creating narratives is an important challenge in NLG (Reiter et al., 2008), and its probably even more important in explanations. Which leads to my third challenge

- *Narrative Challenge:* How can we present the reasoning done by a numerical non-symbolic model, especially one which holistically combines many data elements (e.g., regression and Bayesian models) as a narrative, with key messages linked by causal or argumentative relations?

---

[1]https://shouldiscreen.com/English/lung-cancer-risk-calculator

5 Communicating Uncertainty and Data Quality

People like to think in terms of black and white, yes or no; we are notoriously bad at dealing with probabilities (Kahneman, 2011). One challenge which has received a lot of attention is communicating risk (Berry, 2004; Lundgren and McMakin, 2018); despite all of this attention, it is still a struggle to get people to understand what a 13% risk (for example) really means. Which is a shame, because effective communication of risk in an explanation could really increase scrutability and trust.

Another factor which is important but has received less attention than risk is communicating data quality issues. If we train an AI system on a data set, then any biases in the data set will be reflected in the system's output, For example, if we train a model for predicting lung cancer risks purely on data from Americans, then that model may be substantially less accurate if it is used on people from very different cultures. For instance, few Americans grow up malnourished or in hyper-polluted environments; hence a cancer-prediction model developed on Americans may not accurately estimate risks for a resident of Delhi (one of the most polluted city in the world) who has been malnourished most of her life. Any explanation produced in such circumstances should highlight training bias and any other factors which reduce accuracy.

Similarly, models (regardless of how they are built) may produce inaccurate results if the input data is incomplete or incorrect. For example, suppose someone does not know whether he has a family history of lung cancer (perhaps he is adopted, and has no contact with his birth parents). A lot of AI models are designed to be robust in such cases and still produce an answer; however, their accuracy and reliability may be diminished. In such cases, I think explanations which are scrutable and trustworthy need to highlight this fact, so the user can take this reduced accuracy into consideration when deciding what to do.

There has not been much previous research in data quality in NLG (one exception is Inglis et al. (2017)), which is a shame, because data quality can impact many data-to-text applications, not just explanations. But this does lead to my fourth challenge

- *Communicating Data Quality Challenge:* How can we communicate to users that the accuracy of an AI system is impacted either by the nature of its training data, or by incomplete or incorrect input data?

Of course, communicating uncertainty in the sense of probabilities and risks is also a challenge for both NLG in general and explanations specifically!

6 Conclusion

If we want to produce explanations of AI reasoning in English or other human languages, then we will do a better job if we address the key natural language generation issues of evaluation, user-appropriateness, narrative, and communication of uncertainty and data quality. I have in this paper highlighted four specific challenges within this areas which I think are very important in generating good explanations:

- *Evaluation:* Develop "cheap but reliable" ways of estimating scrutability, trust, etc.

- *Vague Language:* Develop good models for the use of vague language in explanations.

- *Narrative:* Develop algorithms for creating narrative explanations.

- *Data Quality:* Develop techniques to let users know how results are influenced by data issues.

All of these are generic NLG challenges which are important across the board in NLG, not just in explainable AI.

References

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.

Dianne Berry. 2004. *Risk, communication and health psychology*. McGraw-Hill Education (UK).

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.

Or Biran and Kathleen R McKeown. 2017. Human-centric justification of machine learning predictions. In *IJCAI*, pages 1461–1467.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

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.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005-2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60.

James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. 2012. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial intelligence in medicine*, 56(3):157–172.

Stephanie Inglis, Ehud Reiter, and Somayajulu Sripada. 2017. Textually summarising incomplete data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 228–232.

Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

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

Regina E Lundgren and Andrea H McMakin. 2018. *Risk communication: A handbook for communicating environmental, safety, and health risks*. John Wiley & Sons.

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21.

NCBPDeepwaterHorizonSpill. 2011. *Deep water: the gulf oil disaster and the future of offshore drilling: report to the president*. US Government Printing Office. Available at https://www.govinfo.gov/app/details/GPO-OILCOMMISSION.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Cecile Paris. 2015. *User modelling in text generation*. Bloomsbury Publishing.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Ehud Reiter, Albert Gatt, François Portet, and Marian Van Der Meulen. 2008. The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 147–156. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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. ACM.

Statoil. 1993. Discovery evaluation report: Theta vest structure. Available from https://www.equinor.com/en/news/14jun2018-disclosing-volve-data.html.

Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):399–439.

Kees Van Deemter. 2012. *Not exactly: In praise of vagueness*. Oxford University Press.

M.A. Walker, S.J. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, and G. Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.