NL4XAI 2019

1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence

Proceedings of the Workshop

October 29, 2019
Tokyo, Japan

Endorsed by SIGGEN.

Supported by NL4XAI project, which has received funding by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621.

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-950737-70-3

Introduction

Welcome to the Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)!

In the era of the Internet of Things and Big Data, Artificial Intelligence (AI) techniques allow us to automatically extract knowledge from data. This workshop focuses on the automatic generation of interactive explanations in natural language (NL), as humans naturally do, and as a complement to visualization tools. NL technologies, both NL Generation (NLG) and NL Processing (NLP) techniques, are expected to enhance knowledge extraction and representation through human-machine interaction (HMI). As remarked in the last challenge stated by the USA Defense Advanced Research Projects Agency (DARPA), "even though current AI systems offer many benefits in many applications, their effectiveness is limited by a lack of explanation ability when interacting with humans". Accordingly, users without a strong background on AI, require a new generation of Explainable AI systems. They are expected to naturally interact with humans, thus providing comprehensible explanations of decisions automatically made. The ultimate goal is building trustworthy AI that is beneficial to people through fairness, transparency and explainability. To achieve it, not only technical but also ethical and legal issues must be carefully considered.

The workshop will be held as part of the International Conference on Natural Language Generation (INLG2019), which is supported by the Special Interest Group on NLG of the Association for Computational Linguistics. INLG 2019 is to be held in Tokyo (Japan), 29 October - 1 November, 2019. This is the first of a series of workshops to be organized in the next years in the context of the European project NL4XAI (https://nl4xai.eu/).

This half-day workshop goes a step ahead of the workshop 2IS&NLG that we co-organized with Mariët Theune at INLG2018. We have narrowed the workshop topic to become a specialized event on Explainable AI. In this sense, the workshop follows the line started with the workshop XCI at INLG2017. Moreover, this workshop follows a series of thematic special sessions in international conferences such as Advances on Explainable AI at FUZZ-IEEE2019, Advances on Explainable AI at IPMU2018, and many other related sessions at IFSA-EUSFLAT 2009, ISDA 2009, WCCI 2010, WILF 2011, ESTYLF 2012, WCCI 2012, EUSFLAT 2013, IFSA-EUSFLAT2015, and FUZZ-IEEE2017.

Thus, the aim of this workshop is to provide a forum to disseminate and discuss recent advances on Explainable AI. We expect to identify challenges and explore potential transfer opportunities between related fields, generating synergy and symbiotic collaborations in the context of Explainable AI, HMI and Language Generation.

We received 10 submissions (8 regular papers and 2 demos). 4 regular submissions were accepted after a double blind peer review, whereas 2 demos have been included in the program. In addition, NL4XAI 2019 included an invited talk by Dr. Saad Mahamood (*trivago*) to talk about the potential of XAI within industry; and a round table to strengthen the open discussion on the challenges, involving Claire Gardent, Albert Gatt, Ehud Reiter and Jose M. Alonso as panelists.

We would like to thank to all authors for submitting their contributions to our workshop. We thank the program committee members for their work at reviewing the papers and their support during the organization.

Jose M. Alonso and Alejandro Catala
NL4XAI 2019 Organizers

Workshop Organizers:

Jose M. Alonso
Alejandro Catala

Program Committee:

Jose M. Alonso, CiTIUS, University of Santiago de Compostela
Alberto Bugarin, CiTIUS, University of Santiago de Compostela
Katarzyna Budzynska, Institute of Philosophy and Sociology of the Polish Academy of Sciences
Alejandro Catala, CiTIUS, University of Santiago de Compostela
Claire Gardent, CNRS/LORIA, Nancy
Albert Gatt, University of Malta
Dirk Heylen, Human Media Interaction, University of Twente
Corrado Mencar, University of Bari "A. Moro"
Simon Mille, Universitat Pompeu Fabra
Martin Pereira-Fariña, Departamento de Filosofia e Antropoloxıa, University of Santiago de Compostela
Alejandro Ramos-Soto, University of Santiago de Compostela
Chris Reed, Center for Argument Technology, University of Dundee
Ehud Reiter, University of Aberdeen, Arria NLG plc.
Carles Sierra, Institute of Research on Artificial Intelligence (IIIA), Spanish National Research Council (CSIC)
Mariët Theune, Human Media Interaction, University of Twente
Nava Tintarev, Technische University of Delft
Hitoshi Yano, Minsait, INDRA

Invited Speaker:

Saad Mahamood, trivago N.V.

Panelists:

Ehud Reiter (Moderator)
Claire Gardent
Albert Gatt
Jose M. Alonso

Table of Contents

Workshop Program

Tuesday, October 29, 2019

10:00–10:10 Welcome and brief presentation of NL4XAI project

10:10–11:00 Keynote Talk (Saad Mahamood)

11:00–11:30 Oral Presentations I

11:30–11:50 Coffee/Tea Break

11:50–12:20 Oral Presentations II

12:20–12:40 Demo session

12:40–13:20 Round Table

13:20–13:30 Concluding remarks and farewell

The NL4XAI @ INLG 2019 workshop will take place at the National Museum of Emerging Science and Innovation (Miraikan) in Tokyo, Japan, on Tuesday 29th October. Floor 7F, room Saturn.

Invited talk

Explainable Artificial Intelligence and its potential within Industry

Saad Mahamood
saad.mahamood@trivago.com

trivago N.V., Düsseldorf, Germany

Abstract

The age of Big Data has enabled the creation of artificial intelligence solutions that has allowed systems to better respond to their users requests and needs. Applications such as recommender systems, automated content generation systems, etc. are increasingly leveraging such large amounts of data to make better informed decisions about how to tailor their output appropriately. However, the opaqueness of these AI systems in how they derive their decisions or outputs has led to an increasing call for transparency with increasing concerns for the potential of bias to occur in areas such as finance and criminal law. The culmination of these calls have lead to tentative legislative steps. For example, the "Right to explanation" as part of the recently enacted European Union's General Data Protection Regulation.

Natural Language Generation (NLG) has been used in successfully in many data-to-text applications allowing users to gain insights from their data sets. Whilst NLG technology has a strong role to play in generating explanations for AI models there still remains inherit challenges in developing and deploying text generation systems within a commercial context.

In this talk I will explore the role and potential that Natural Language Explainable AI can have within *trivago* and the wider industry. *trivago* is a leading accommodation meta-search engine that enables users find the right hotel or apartment at the right price. In particular, this talk will describe the work we have done to apply natural language solutions within *trivago* and the challenges of applying AI solutions from a commercial perspective. Finally, this talk will also explore the potential applications of where explainable AI approaches could be used within *trivago*.

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 1–2,
Tokyo, Japan, October 29 2019.

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 using 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",

> 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

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.

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 3–7,
Tokyo, Japan, October 29 2019.

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.

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

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?

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.

Acknowledgments

This paper started off as a (much shorter) blog https://ehudreiter.com/2019/07/19/nlg-and-explainable-ai/. My thanks to the people who commented on this blog, as well as the anonymous reviewers, the members of the Aberdeen CLAN research group, the members of the Explaining the Outcomes of Complex Models project at Monash, and the members of the NL4XAI research project, all of whom gave me excellent feedback and suggestions. My thanks also to Prof René van der Wal for his help in the experiment mentioned in section 3.

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.

A Survey of Explainable AI Terminology

Miruna A. Clinciu and Helen F. Hastie
Edinburgh Centre for Robotics
Heriot-Watt University, Edinburgh, EH14 4AS, UK
{mc191, H.Hastie}@hw.ac.uk

Abstract

The field of Explainable Artificial Intelligence attempts to solve the problem of algorithmic opacity. Many terms and notions have been introduced recently to define Explainable AI, however, these terms seem to be used interchangeably, which is leading to confusion in this rapidly expanding field. As a solution to overcome this problem, we present an analysis of the existing research literature and examine how key terms, such as *transparency*, *intelligibility*, *interpretability*, and *explainability* are referred to and in what context. This paper, thus, moves towards a standard terminology for Explainable AI.

***Keywords*—** Explainable AI, black-box, NLG, Theoretical Issues, Transparency, Intelligibility, Interpretability, Explainability

1 Introduction

In recent years, there has been an increased interest in the field of Explainable Artificial Intelligence (XAI). However, there is clear evidence from the literature that there are a variety of terms being used interchangeably such as *transparency*, *intelligibility*, *interpretability*, and *explainability*, which is leading to confusion. Establishing a set of standard terms to be used by the community will become increasingly important as XAI is mandated by regulation, such as the GDPR and as standards start to appear such as the IEEE standard in transparency (P7001). This paper works towards this goal.

Explainable Artificial Intelligence is not a new area of research and the term **explainable** has existed since the mid-1970s (Moore and Swartout, 1988). However, XAI has come to the forefront in recent times due to the advent of deep machine learning and the lack of transparency of "black-box" models. We introduce below, some descriptions of XAI collected from the literature:

- "Explainable AI can present the user with an easily understood chain of reasoning from the user's order, through the AI's knowledge and inference, to the resulting behaviour" (van Lent et al., 2004).
- "XAI is a research field that aims to make AI systems results more understandable to humans" (Adadi and Berrada, 2018).

Thus, we conclude that XAI is a research field that focuses on giving AI decision-making models the ability to be easily understood by humans. Natural language is an intuitive way to provide such Explainable AI systems. Furthermore, XAI will be key for both expert and non-expert users to enable them to have a deeper understanding and the appropriate level of trust, which will hopefully lead to increased adoption of this vital technology.

This paper firstly examines the various notions that are frequently used in the field of Explainable Artificial Intelligence in Section 2 and attempts to organise them diagrammatically. We then discuss these terms with respect to Natural Language Generation in Section 3 and provide conclusions.

2 Terminology

In this section, we examine four key terms found frequently in the literature for describing various techniques for XAI. These terms are illustrated in Figure 1, where we organise them as a Venn diagram that describes how a *transparent* AI system has several facets, which include *intelligibility*, *explainability*, and *interpretability*. Below, we discuss how *intelligibility* can be discussed in terms of *explainability* and/or *interpretability*. For each of these terms, we present the dictionary definitions extracted from modern and notable English dictionaries, quotes from the literature presented in tables and discuss how they support the proposed structure given in Figure 1. In every table, we emphasise related words and context, in order

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 8–13,
Tokyo, Japan, October 29 2019.

to connect ideas and build up coherent relationships within the text.

In this paper, the first phase of the selection criteria of publications was defined by the relevance of the paper and related key words. The second phase was performed manually by choosing the papers that define or describe the meaning of the specified terms or examine those terms for ways in which they are different, alike, or related to each other.

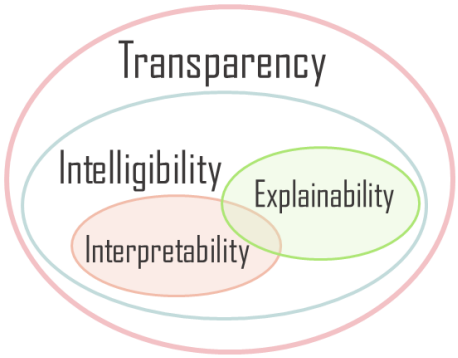

Figure 1: A Venn Diagram of the relationship between frequently used terms, that offers a representation of the authors' interpretation for the field, excluding post-hoc interpretation.

Transparency

Dictionary definitions: The word "*transparent*" refers to something that is "clear and easy to understand" (Cambridge Dictionary, 2019d); or "easily seen through, recognized, understood, detected; manifest, evident, obvious, clear" (Oxford English Dictionary, 2019d); or "language or information that is transparent is clear and easy to understand" (The Longman Dictionary of Contemporary English, 2019c).

Conversely, an opaque AI system is a system with the lowest level of transparency, known as a "**black-box**" model. A similar definition is given by Tomsett et al. (2018) in Table 1.

Tintarev and Masthoff (2007) state that *transparency* "explains how the system works" and it is considered one of the possible explanation facilities that could influence good recommendations in recommender systems.

In the research paper by Cramer et al. (2008), *transparency* aims to increase understanding and entails offering the user insight as to how a system works, for example, by offering explanations for system choices and behaviour.

"**Transparency** clearly describing the model structure, equations, parameter values, and assumptions to enable interested parties to understand the model" (Briggs et al., 2012).
Tomsett et al. (2018) defined **transparency** as a "level to which a system provides information about its internal workings or structure" and both "**explainability** and **transparency** are important for improving creator-interpretability".
"Informally, **transparency** is the opposite of opacity or blackbox-ness. It connotes some sense of understanding the mechanism by which the model works. We consider **transparency** at the level of the model (simulatability), at the level of individual components (e.g. parameters) (decomposability), and at the level of the training algorithm (algorithmic transparency)" (Lipton, 2016).

Table 1: Various notions of Transparency presented in recent research papers

Intelligibility

Dictionary definitions: An "*intelligible*" system should be "clear enough to be understood" according to Cambridge Dictionary (2019b); or "capable of being understood; comprehensible" (Oxford English Dictionary, 2019b); or "easily understood" (The Longman Dictionary of Contemporary English, 2019d).

The concept of *intelligibility* was defined by Bellotti and Edwards (2001) from the perspective of "context-aware systems that seek to act upon what they infer about the context must be able to represent to their users what they know, how they know it, and what they are doing about it" (Bellotti and Edwards, 2001).

As illustrated in Table 2, it is challenging to define how intelligible AI systems could be designed, as they would need to communicate very complex computational processes to various types of users (Weld and Bansal, 2018). Per the Venn diagram in Figure 1, we consider that an AI system could become intelligible in a number of ways, but also through *explanations* (e.g. in natural language) and/or *interpretations*. We discuss both of these in turn below.

"It remains remarkably hard to specify what makes a system **intelligible**; The key challenge for designing **intelligible AI** is communicating a complex computational process to a human. Specifically, we say that a model is **intelligible** to the degree that a human user can predict how a change to a feature" (Weld and Bansal, 2018).
"**Intelligibility** can help expose the inner workings and inputs of context-aware applications that tend to be opaque to users due to their implicit sensing and actions" (Lim and Dey, 2009).

Table 2: Various notions of Intelligibility presented in recent research papers

Interpretability

Dictionary Definitions: According to Cambridge Dictionary (2019c), the word "*interpret*" definition is "to decide what the intended meaning of something is"; or "to expound the meaning of (something abstruse or mysterious); to render (words, writings, an author, etc.) clear or explicit; to elucidate; to explain" (Oxford English Dictionary, 2019c); or "to explain the meaning of something" (The Longman Dictionary of Contemporary English, 2019b).

Considering a "black-box" model, we will try to understand how users and developers could define the model *interpretability*. A variety of definitions of the term *interpretability* have been suggested in recent research papers, as presented in Table 3.

Various techniques have been used to give insights into an AI model through interpretations, such as Feature Selection Techniques (Kim et al., 2015), Shapley Values (Sundararajan and Najmi, 2019); the interpretation of the AI model interpretation e.g. Hybrid AI models (Wang and Lin, 2019), by combining interpretable models with opaque models, and output interpretation (e.g. Evaluation Metrics Interpretation (Mohseni et al., 2018), and Visualisation Techniques Interpretation (Samek et al., 2017; Choo and Liu, 2018)). Thus in our model in Figure 1, we define *interpretability* as intersecting with *explainability* as some models may be interpretable without needing explanations.

"In model-agnostic **interpretability**, the model is treated as a black-box. **Interpretable models** may also be more desirable when interpretability is much more important than accuracy, or when interpretable models trained on a small number of carefully engineered features are as accurate as black-box models". (Ribeiro et al., 2016)
"An **explanation** can be evaluated in two ways: according to its **interpretability**, and according to its completeness" (Gilpin et al., 2018).
"We define **interpretable** machine learning as the use of machine-learning models for the extraction of relevant knowledge about domain relationships contained in data..." (Murdoch et al., 2019).

Table 3: Various notions of Interpretability presented in recent research papers

Explainability

Dictionary Definitions: For the word "*explain*" were extracted the following definitions: "to make something clear or easy to understand by describing or giving information about it" Cambridge Dictionary (2019a); or "to provide an explanation for something. to make plain or intelligible" (Oxford English Dictionary, 2019a); or "to tell someone about something in a way that is clear or easy to understand. to give a reason for something or to be a reason for something" (The Longman Dictionary of Contemporary English, 2019a).

Per these definitions, providing explanations is about improving the user's mental model of how a system works. Ribera and Lapedriza (2019) consider that we do not have a concrete definition for *explanation* in the literature. However, according to these authors, every definition relates "explanations with "why" questions or causality reasonings". Given the nature of the explanations, Ribera and Lapedriza (2019) proposed to categorise the explainees in three main groups, based on their goals, background, and relationship with the product, namely: developers and AI researchers, domain experts, and lay users. Various types of explanations have been presented in the literature such as "why" and "why not" (Kulesza et al., 2013) or Adadi and Berrada (2018)'s four types of explanations that are used to "justify, control, discover and improve". While it is out of scope

to go into detail here, what is clear is that in most uses of the term *explainability*, it means providing a way to improve the understanding of the user, whomever they may be.

"**Explanation** is considered closely related to the concept of **interpretability**" (Biran and Cotton, 2017).
"**Transparent** design: model is inherently **interpretable** (globally or locally)" (Lucic et al., 2019).
"I equate **interpretability** with **explainability**" (Miller, 2018).
"Systems are **interpretable** if their operations can be understood by a human, either through **introspection** or through a **produced explanation**" (Biran and Cotton, 2017).
In the paper (Poursabzi-Sangdeh et al., 2018), **interpretability** is defined as something "that cannot be manipulated or measured, and could be defined by people, not algorithms".

Table 4: Various notions of Explainability presented in recent research papers

3 The Role of NLG in XAI

An intuitive medium to provide such explanations is through natural language. The human-like capability of Natural Language Generation (NLG) has the potential to increase the intelligibility of an AI system and enable a system to provide explanations that are tailored to the end-user (Chiyah Garcia et al., 2019).

One can draw an analogy between natural language generation of explanations and Lacave and Diez's model of explanation generation for expert systems (Lacave and Díez, 2002); or Reiter and Dale's NLG pipeline (Reiter and Dale, 2000) with stages for determining "what" to say in an explanation (content selection) and "how" to say it (surface realisation). Lacave and Diez's model also emphasises the importance of adapting to the user, which is also a focus area in NLG (e.g. adapting styles (Dethlefs et al., 2014)).

Other studies have looked at agents and robots providing a rationalisation of their behaviour (Ehsan et al., 2018) by providing a running commentary in language. Whilst this is not necessarily how humans behave, it is beneficial to be able to provide such *rationalisation*, especially in the face of unusual behaviour and, again, natural language is one way to do this. Defined as a process of producing an explanation for an agent or system behavior as if a human had performed the behaviour, *AI rationalisation* has multiple advantages to be taken into consideration: "naturally accessible and intuitive to humans, especially non-experts, could increase the satisfaction, confidence, rapport, and willingness to use autonomous systems and could offer real-time response" (Ehsan et al., 2018).

4 Conclusions and Future work

In this paper, we introduced various terms that could be found in the field of Explainable AI and their concrete definition. In Figure 1, we have attempted to define the relationship between the main terms that define Explainable AI. Intelligibility could be achieved through explanations and interpretations, where the type of user, their background, goal and current mental model are taken into consideration.

As mentioned previously, *interpretability* is defined as a concept close to *explainability* (Biran and Cotton, 2017). Our Venn diagram given in Figure 1 illustrates that transparent systems could be, by their nature interpretable, without providing explanations and that the activities of interpreting a model and explaining why a system behaves the way it does are fundamentally different. We posit, therefore, that the field moving forward should be wary of using such terms interchangeably. Natural Language Generation will be key to providing explanations, and rationalisation is one approach that we have discussed here.

Evaluation of NLG is challenging area (Hastie and Belz, 2014) with objective measures such as BLEU being shown not to reflect human ratings (Liu et al., 2016). How natural language explanations are evaluated will likely be based on, in the near term at least, subjective measures that try to evaluate an explanation in terms of whether it improves a system's *intelligibility*, *interpretability* and *transparency* along with other typical metrics related to the quality and clarity of the language used (Curry et al., 2017).

In future work, it would be advisable to perform empirical analysis of research papers related to the various terms and notions introduced here and continuously being added into the field of XAI.

Acknowledgements

The authors gratefully acknowledge the support of Dr. Inês Cecilio, Prof. Mike Chantler, and Dr. Vaishak Belle. This research was funded by Schlumberger Cambridge Research Centre Doctoral programme.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: Human considerations in context-aware systems. *Human-Computer Interaction*, 16(2-4):193–212.

Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. In *Proceedings of the 1st Workshop on Explainable Artificial Intelligence, IJCAI 2017*.

Andrew H. Briggs, Milton C. Weinstein, Elisabeth A. L. Fenwick, Jonathan Karnon, Mark J. Sculpher, and A. David Paltiel. 2012. Model parameter estimation and uncertainty: A report of the ispor-smdm modeling good research practices task force-6. *Value in Health*, 15(6):835–842.

Cambridge Dictionary. 2019a. *Explain*. Cambridge University Press. Accessed on 2019-08-25.

Cambridge Dictionary. 2019b. *Intelligible*. Cambridge University Press. Accessed on 2019-08-25.

Cambridge Dictionary. 2019c. *Interpret*. Cambridge University Press. Accessed on 2019-08-25.

Cambridge Dictionary. 2019d. *Transparent*. Cambridge University Press. Accessed on 2019-08-25.

Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2019. Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models. In *Proceedings of the International Natural Language Generation (INLG)*.

J. Choo and S. Liu. 2018. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, 38(4):84–92.

Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455.

Amanda Cercas Curry, Helen Hastie, and Verena Rieser. 2017. A review of evaluation techniques for social dialogue systems. In *ISIAA 2017 - Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, Co-located with ICMI 2017*, pages 25–26. Association for Computing Machinery, Inc.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 702–711. Association for Computational Linguistics (ACL).

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 81–87. Association for Computing Machinery, Inc.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of the 5th International Conference on Data Science and Advanced Analytics (DSAA) 2018 IEEE*, pages 80–89. IEEE.

Helen Hastie and Anja Belz. 2014. A comparative evaluation methodology for nlg in interactive systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Been Kim, Julie A Shah, and Finale Doshi-Velez. 2015. Mind the gap: A generative approach to interpretable feature selection and extraction. In *Proceedings of the Twenty-ninth Conference on Neural Information Processing Systems, NeurIPS 2015*, pages 2260–2268. Curran Associates, Inc.

T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. Wong. 2013. Too much, too little, or just right? ways explanations impact end users' mental models. In *Proceedings of the 2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10. IEEE.

Carmen Lacave and Francisco J. Díez. 2002. A Review of Explanation Methods for Bayesian Networks. *The Knowledge Engineering Review*, 17(2):107–127.

Michael van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence*, IAAI'04, pages 900–907. AAAI Press.

Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*, UbiComp '09, pages 195–204, New York, NY, USA. ACM.

Zachary Chase Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, page 13.

Ana Lucic, Hinda Haned, and Maarten de Rijke. 2019. Contrastive explanations for large errors in retail forecasting predictions through monte carlo simulations. *arXiv preprint arXiv:1908.00085*.

Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv preprint arXiv:1706.07269*.

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*, abs/1811.11839.

J.D. Moore and W.R. Swartout. 1988. *Explanation in Expert Systems: A Survey*. Number no. 228 in Explanation in Expert Systems: A Survey. University of Southern California, Information Sciences Institute.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.

Oxford English Dictionary. 2019a. *explain, v.* Oxford University Press. Accessed on 2019-11-10.

Oxford English Dictionary. 2019b. *intelligible, adj. (and n.)*. Oxford University Press. Accessed on 2019-11-10.

Oxford English Dictionary. 2019c. *interpret, v.* Oxford University Press. Accessed on 2019-11-10.

Oxford English Dictionary. 2019d. *transparent, adj. (and n.)*. Oxford University Press. Accessed on 2019-11-10.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *Proceedings of the CEUR Workshop*, volume 2327. CEUR-WS.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Mller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, 1:1–10.

Mukund Sundararajan and Amir Najmi. 2019. The many shapley values for model explanation. *arXiv preprint arXiv:1908.08474*, abs/1908.08474.

The Longman Dictionary of Contemporary English. 2019a. *explain*. Pearson Longman. Accessed on 2019-11-10.

The Longman Dictionary of Contemporary English. 2019b. *interpret, v.* Pearson Longman. Accessed on 2019-11-10.

The Longman Dictionary of Contemporary English. 2019c. *transparent, adj. (and n.)*. Pearson Longman. Accessed on 2019-11-10.

The Longman Dictionary of Contemporary English. 2019d. *transparent, adj. (and n.)*. Pearson Longman. Accessed on 2019-11-10.

Nava Tintarev and Judith Masthoff. 2007. Effective explanations of recommendations: User-centered design. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, pages 153–156, New York, NY, USA. ACM.

Richard Tomsett, Dave Braines, Dan Harborne, Alun D. Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, abs/1806.07552.

Tong Wang and Qihang Lin. 2019. Hybrid predictive model: When an interpretable model collaborates with a black-box model. *arXiv preprint arXiv:1905.04241*.

Daniel S. Weld and Gagan Bansal. 2018. Intelligible artificial intelligence. *arXiv preprint arXiv:1803.04263*.

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.

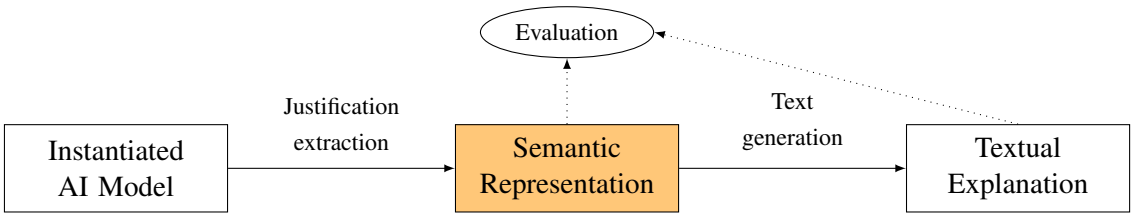

Figure 1: XAI architecture proposal to produce and evaluate explanations

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.

Paving the way towards counterfactual generation in argumentative conversational agents

Ilia Stepin, Alejandro Catalá, Jose M. Alonso
Centro Singular de Investigación en
Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago
de Compostela, Spain
{ilia.stepin,alejandro.catala,
josemaria.alonso.moral}@usc.es

Martín Pereira-Fariña
Departamento de Filosofía
e Antropoloxía,
Universidade de Santiago
de Compostela, Spain
martin.pereira@usc.es

Abstract

Counterfactual explanations present an effective way to interpret predictions of black-box machine learning algorithms. Whereas there is a significant body of research on counterfactual reasoning in philosophy and theoretical computer science, little attention has been paid to counterfactuals in regard to their explanatory capacity. In this paper, we review methods of argumentation theory and natural language generation that counterfactual explanation generation could benefit from most and discuss prospective directions for further research on counterfactual generation in explainable Artificial Intelligence.

1 Introduction

Automatic decision-making systems using black-box machine learning (ML) algorithms are now widely used in various complex domains from legislation (Greenleaf et al., 2018) to health care (Gargeya and Leng, 2017). However, such systems cannot be trusted blindly as their output often comes unexplained to end users (Rudin, 2018). As a result, there exists a lack of confidence in such automatic decisions caused by a low degree of their interpretability (Ribeiro et al., 2016).

The need for intelligent systems to explain their decisions has driven a decent amount of research in the past decades (Biran and Cotton, 2017). However, advances in social sciences impose novel challenges on explainable agents. For example, recent findings from cognitive science testify that the key feature of explanations is their contrastiveness (Miller, 2019), that is the ability to reflect on alternative scenarios of actually happened events. Whereas little research has been performed on generation of such counterfactual explanations, we believe that enabling virtual assistants and recommendation systems with the ability to generate them should increase greatly their acceptance among end users.

In this paper, we briefly review prospective methods for addressing the problem of counterfactual explanation generation. Subsequently, we aim to further shape the line of research devoted to counterfactual analysis for explainable Artificial Intelligence (AI) by pointing to the existing field-specific theoretical foundations and potential directions of its algorithmic design. As a result, this work supports a discussion on prospective methods for argumentative conversational agent development.

The rest of the manuscript is organised as follows. Section 2 inspects definitions of a counterfactual explanation and reviews existing generation approaches to counterfactual explanations. Section 3 describes the most prominent formal argumentation frameworks as a theoretical basis for counterfactual analysis. Section 4 discusses the classification of and recent advances in developing argumentative conversational agents in the context of counterfactual generator implementation. Finally, we conclude with outlining open challenges relevant for counterfactual explanation generation in section 5.

2 Counterfactual explanations

Explanations are argued to be contrastive (Miller, 2019). According to Miller, people are not satisfied with mere direct explanations in form of causal relations between the antecedent and consequent but also require to know why an alternative (or opposing) event could not have happened. Furthermore, Pearl and Mackenzie (2018) argue that it is the ability to produce such contrastive statements, referred to as *counterfactuals*, that lies on top of human reasoning.

In ML, a counterfactual explanation describes

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 20–25,
Tokyo, Japan, October 29 2019.

an alternative (hypothesised) situation which is as similar as possible to the original event in terms of its feature values while having a different outcome prediction ("the closest possible world") (Molnar, 2019). When searching for a suitable counterfactual explanation, the distance between a given piece of factual information and its counterpart is to be minimised while the outcome is different so that the counterfactual presumes only the most relevant alterations to the original fact. In addition, counterfactuals capture contextual information as they describe "a dependency on the external facts that led to a decision" (Wachter et al., 2018). As a result, explanations supported by counterfactuals are likely to gain acceptability by end users.

While the general understanding of the concept of counterfactuals is shared among researchers, there exist several interpretations of this phenomenon. As counterfactuals are generally assumed to have a clear connection with causation (Pearl and Mackenzie, 2018), they are often viewed as non-observable potential outcomes that would have happened in the absence of the cause (Shadish et al., 2002). In terms of causality, they are informally defined as conditional statements in the form: "If event X had not occurred, event Y would not have occurred" (Lewis, 1973). However, Wachter et al. (2018) propose a causation-free definition of an unconditional counterfactual statement based on the idea of subject's disbelief in a given hypothetical situation. On the other hand, counterfactuals are also sometimes referred to as "conditional connectives" in conditional logic (Besnard et al., 2013).

In recent years, there have been several attempts to approach the problem of counterfactual explanation generation. Wachter et al. (2018) suggested an approach for calculating counterfactuals based on the use of the Manhattan distance. Sokol and Flach (2018) adopted this approach to implement a counterfactual explanation generator for a decision tree-based AI system. In addition, Hendricks et al. (2018) proposed a model where candidate counterfactual pieces of evidence are selected from a set of all the noun-phrases of the corresponding textual descriptions of input images. Such evidence is then verified to be absent in the original image so that it can be used in the output counterfactual explanation. A rule-based system is then used to generate fluent negated explanations. Later, Birch et al. (2019) introduced an arbitrated dispute tree model arguing that the explanations generated by their model are indeed contrastive in accordance with the principles proposed by Miller (2019) as opposite outcomes are presented for all cases. Furthermore, the corresponding features and stages are explicitly found for cases opposing to the focus case.

As has been shown above, the problem of counterfactual explanation generation is concerned with several topics from philosophy, (computational) linguistics, and AI. While this leaves room for developing novel synergistic methods and algorithms that would combine insights from all the relevant fields, potential challenges when developing such tools are multiplied. For example, the fact that certain types of counterfactual explanations are preferred over their counterparts (Byrne, 2019) places further restrictions on newly developed frameworks as in designing heuristics for reducing the search space of the most relevant counterfactual explanations in accordance with such additional restrictive criteria.

In conclusion, counterfactual explanations are likely to enrich conversational interfaces of any system to be considered explainable. However, counterfactuals produced directly from ML algorithm predictions show a lack of coherence and appear unreliable from the ethical point of view (Kusner et al., 2017). Moreover, they usually do not involve a user in an extensive dialogic interaction, which makes them self-explanatory only in a limited number of cases. Therefore, we hypothesise that going deeper with their formalization is likely to overcome these weaknesses.

3 Formal argumentation

Formal argumentation (Baroni et al., 2018) provides practitioners with a natural form of counterfactual explanation formalization. Indeed, argumentation is claimed to mimic human reasoning (Cerutti et al., 2014). As such, it offers a set of tools that have become widely applicable to interpreting the output of ML algorithms. Formal argumentation embraces a wide range of theoretical frameworks from *argumentation schemes* (Walton et al., 2008) to *dialogue games* (Carlson, 1985), among others. In this paper, we focus on *abstract argumentation* (AA) frameworks as a prospective theoretical basis for counterfactual explanation generation.

While disregarding the internal structure of ar-

guments, AA frameworks primarily deal with relations between arguments. The AA framework introduced in Dung (1995) is a pioneering theoretical framework, which has become well known. This AA framework is a directed graph (also referred to as "argument graph") formally defined as a pair $AA = (A, R)$ where A is a set of arguments, $R \subseteq A \times A$ being a set of binary *attack* relations between pairs of arguments $(a, b) \in R$. In these settings, argument a is assumed to attack argument b. The acceptability of arguments is defined through numerous semantics in form of extensions over a conflict-free set of arguments, which is defined as a subset of all arguments that do not attack each other.

Due to its seeming simplicity, Dung's framework only presents the very basic argumentative constructs. Indeed, a number of extensions address this handicap. For example, some models attempt to extend the original Dung's argumentation framework by refining the concept of attacks between arguments allowing attack-to-attack relations (Modgil, 2007; Baroni et al., 2011). In contrast, a significant body of research aims to complement the nature of relations between arguments by incorporating supportive relations (Verheij, 2002; Amgoud et al., 2008).

It is worth noting that variants of AA have already been employed to address the problem of explanation generation. For example, Amgoud and Serrurier (2008) use the AA framework to resolve a binary classification task and motivate the outcome with arguments constructed, subsequently compared against each other, and ranked according to their strength. Šešelja and Straßer (2013) augment AA with explanatory features for scientific debate modelling. However, none of these works embodies counterfactual explanations.

Dung et al. (2009) proposed a conceptually novel instance of the AA framework which is known as the assumption-based argumentation (ABA) framework. Thus, ABA operates on a set of assumptions deducted via inference rules and reconsiders attack relations defined now as contraries to assumptions supporting the original argument. Following this approach, Zhong et al. (2019) implements an ABA multi-attribute explainable decision model that generates textual explanations on the basis of dispute trees. Notice that this model is claimed to be an argumentation-based framework to generate textual explanations for decision-making models. Nevertheless, while justifying why a particular decision is preferred over its counterpart, the model does not offer counterfactual explanations for rejected decisions.

Despite a rising interest towards counterfactual explanation generation in recent years, little work has been done in the direction of applying formal methods (including argumentation) to generation of counterfactual explanations. While most existing counterfactual frameworks make use of elements of causal inference, we find counterfactual statements naturally integrated into conditional logic-based (Besnard et al., 2013) as well as abstract argumentation (Sakama, 2014) frameworks. However, none of these frameworks governs any existing counterfactual explanation generation system so far.

4 Argumentative conversational agents

Argumentative frameworks can be embedded directly into chatbots or conversational agents to interact with end users. In terms of practical implementation, conversational agents are broadly divided into two main groups: retrieval-based and generative agents (Chen et al., 2017). On the one hand, a retrieval-based agent aims to select the most suitable response from the set of predefined responses that it contains given user's inquiry (Rakshit et al., 2017; Bartl and Spanakis, 2017). This kind of agents is based on the use of templates and produces grammatical utterances in all cases. However, such template-based text generators are expensive to develop and maintain due to immense expert labour resources required. On the other hand, generative models can form previously unseen utterances as they are trained from scratch without any templates in store (Li et al., 2016; Shao et al., 2017). Nevertheless, their generic responses limit their applicability to explainable AI problems.

The need for explanability of complex ML-based systems imposes additional requirements on conversational agents. Thus, automatically generated explanations are expected to be convincing enough in order to increase user's confidence in system's predictions with respect to the given task. This is hypothesised to lead to an indispensable shift of attention towards development of argumentative conversational agents (or argumentative dialogue systems) operating on a set of arguments as responses to user's inquiries. Further-

more, such argumentation-based agents are considered to push the boundaries of the present-day conversational agents towards more human-like interaction (Dignum and Bex, 2018). In combination with recent advances in deep learning and reinforcement learning, the use of argumentation as a theoretical basis for conversational agents opens prospects for a new era of generative conversational agents (Rosenfeld and Kraus, 2016; Rach et al., 2019).

Finally, the issue of evaluation of argumentation-based conversational agents merges with those coming directly from the field of natural language generation (NLG) and explainable AI. At present, there is no unifying agreement on a set of evaluation metrics to be used neither within the NLG community (Gatt and Krahmer, 2018) nor within the explainable AI community (Adadi and Berrada, 2018). While common objective (automatic) and subjective (human-oriented surveys) metrics used for NLG evaluation are found in the literature on conversational agents and dialogue systems, novel metrics are regularly introduced for instances of argumentative chatbots (e.g., distinctiveness, as in (Le et al., 2018)) and counterfactual generators (e.g., accuracy with counterfactual text and phrase-error, as in (Hendricks et al., 2018)). Thus, a direct comparison between analogous agents becomes a particularly challenging task. As a possible solution, a combination of subjective and objective metrics is believed to be a reasonable starting point for a discussion on the choice of evaluation techniques. At the same time, automatically generated explanations are expected to be accurate, consistent, and comprehensible. As the perception of these properties is highly subjective, they cannot be measured (and therefore evaluated) directly and require further investigation.

5 Concluding remarks

Our literature review has revised the foundations of current approaches to counterfactual explanation generation. The limitations found call for some potential areas for improvement on the development of explainable AI systems.

First, there is no single definition of a counterfactual explanation. While counterfactuals have various interpretations in the literature, we find it particularly important to suggest a uniform definition that would not only capture all the properties of counterfactual explanations but also allow for designing a universal domain-independent framework for their generation.

Second, existing argumentation-based explanation generation models do not fully solve the problem of counterfactual explanation generation. While some of such models do not offer consistent explanations in textual form, others do not output contrastive explanations. Therefore, a more holistic counterfactual generation framework should be developed to close this gap.

Third, formal argumentation is rarely considered in present-day conversational agents. To the best of our knowledge, such argumentation-based agents do not consider incoming dialogic information received from the direct interaction with the user to contextualise their counterfactual explanations. However, processing such information may help to improve the quality of the offered counterfactual explanations making them more personalised. Therefore, capturing such contextual information presents another noteworthy line of research.

The aforementioned issues, along with others not discussed due to space limitations, show that the generation of counterfactual explanations is a timely but complex problem. In the future, we plan to address these issues by designing an argumentation-based dialogue protocol and developing a conversational agent ready to make use of the protocol to output accurate and consistent counterfactual explanations.

Acknowledgments

Jose M. Alonso is *Ramón y Cajal* Researcher (RYC-2016-19802). This research was also funded 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 and "accreditation 2016-2019, ED431G/08"). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable arti-

ficial intelligence (XAI) . *IEEE Access*, 6:52138–52160.

Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex, and Pierre Livet. 2008. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23(10):1062–1093.

Leila Amgoud and Mathieu Serrurier. 2008. Agents that argue and explain classifications. *Autonomous Agents and Multi-Agent Systems*, 16(2):187–209.

P. Baroni, D. Gabbay, and M. Giacomin. 2018. *Handbook of Formal Argumentation*. College Publications.

Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Giovanni Guida. 2011. AFRA : Argumentation framework with recursive attacks. *International Journal of Approximate Reasoning*, 52(1):19–37.

A. Bartl and G. Spanakis. 2017. A retrieval-based dialogue system utilizing utterance and context embeddings. In *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1120–1125, United States.

Philippe Besnard, Éric Grégoire, and Badran Raddaoui. 2013. A conditional logic-based argumentation framework. In *Scalable Uncertainty Management*, pages 44–56, Berlin, Heidelberg. Springer Berlin Heidelberg.

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence (XAI)*, pages 8–13.

David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. 2019. Explanations by arbitrated argumentative dispute. *Expert Systems With Applications*, 127:141–156.

Ruth M. J. Byrne. 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6276–6282.

Laura Carlson. 1985. *Dialogue Games: An Approach to Discourse Analysis*. Reidel.

Federico Cerutti, Nava Tintarev, and Nir Oren. 2014. Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, ECAI'14, pages 207–212. IOS Press.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19:25–35.

Frank Dignum and Floris Bex. 2018. Creating dialogues using argumentation and social practices. In *Internet Science*, pages 223–235. Springer International Publishing.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357.

Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. 2009. *Assumption-Based Argumentation*, pages 199–218. Springer US, Boston, MA.

Rishab Gargeya and Theodore Leng. 2017. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124:962–969.

Albert Gatt and E.J. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1):65–170.

Graham Greenleaf, Andrew Mowbray, and Philip Chung. 2018. Building sustainable free legal advisory systems: Experiences from the history of AI & law. *Computer Law & Security Review*, 34(2):314–326.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating counterfactual explanations with natural language. In *Proceedings of the Workshop on Human Interpretability in Machine Learning (WHI)*, pages 95–98.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 4069–4079.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130, Brussels, Belgium. Association for Computational Linguistics.

David K. Lewis. 1973. *Counterfactuals*. Blackwell.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

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

Sanjay Modgil. 2007. An abstract theory of argumentation that accommodates defeasible reasoning about preferences. In *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 648–659. Springer Berlin Heidelberg.

Christoph Molnar. 2019. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Leanpub.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*, 1st edition. Basic Books, Inc., New York, NY, USA.

N. Rach, K. Weber, A. Aicher, F. Lingenfelser, E. André, and W. Minker. 2019. Emotion recognition based preference modelling in argumentative dialogue systems. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 838–843.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems (IWSDS)*, pages 45–52.

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, New York, NY, USA.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, volume 285, pages 320–328.

Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. In *Proceedings of the Workshop on Critiquing and Correcting Trends in Machine Learning – 32nd Conference on Neural Information Processing Systems (NIPS)*.

Chiaki Sakama. 2014. Counterfactual reasoning in argumentation frameworks. In *Proceedings of the 5th International Conference on Computational Models of Argument (COMMA)*, pages 385–396.

Dunja Šešelja and Christian Straßer. 2013. Abstract argumentation and explanation applied to scientific debates. *Synthese*, 190(12):2195–2217.

W. R. Shadish, T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2210–2219.

Kacper Sokol and Peter A. Flach. 2018. Conversational explanations of machine learning predictions through class-contrastive counterfactual statements. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5785–5786.

Bart Verheij. 2002. On the existence and multiplicity of extensions in dialectical argumentation. In *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning (NMR)*, pages 416–425.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law Technology*, 31:841–887.

D. Walton, P.C. Reed, C. Reed, and F. Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni. 2019. An explainable multi-attribute decision model based on argumentation. *Expert Systems With Applications*, 117:42–61.

Engaging in Dialogue about an Agent's Norms and Behaviors

Daniel Kasenberg*, Antonio Roque, Ravenna Thielstrom, and **Matthias Scheutz**
Human-Robot Interaction Laboratory
Tufts University
Medford, MA, USA
*dmk@cs.tufts.edu

Abstract

We present a set of capabilities allowing an agent planning with moral and social norms represented in temporal logic to respond to queries about its norms and behaviors in natural language, and for the human user to add and remove norms directly in natural language. The user may also pose hypothetical modifications to the agent's norms and inquire about their effects.

1 Introduction and Related Work

Explainable planning (Fox et al., 2017) emphasizes the need for developing artificial agents which can explain their decisions to humans. Understanding how and why an agent made certain decisions can facilitate human-agent trust (Lomas et al., 2012; Wang et al., 2016; Garcia et al., 2018).

At the same time, the field of *machine ethics* emphsizes developing artificial agents capable of behaving ethically. Malle and Scheutz (2014) have argued that artificial agents ought to obey human moral and social norms (rules that humans both obey and expect others to obey), and to communicate in terms of these norms. Some have argued in favor of using temporal logic to represent agent objectives, including moral and social norms (e.g. Arnold et al., 2017; Camacho and Mcilraith, 2019), in particular arguing that it can capture complex goals while remaining interpretable in a way that other methods (e.g. reinforcement learning) are not. Nevertheless, explaining behavior in terms of temporal logic norms has been little considered (though see Raman et al., 2016).

In this paper we consider an artificial agent planning to maximally satisfy some set of moral and social norms, represented in an object-oriented temporal logic. We present a set of capabilities for such an agent to respond to a human user's queries as well as to commands adding and removing norms, both actually and hypothetically (and thus taking a step toward two-way *model reconciliation* (Chakraborti et al., 2017), in which agent and human grow to better understand each other's models and values).

2 Contribution

Our system enables an agent planning with norms specified in an object-oriented temporal logic called violation enumeration language (VEL) to explain its norms and its behavior to a human user; the user may also directly modify the agent's norms via natural language (both really and hypothetically). While the planner and the system used to generate the (non-NL) can handle a broad subset of VEL statements, our natural language systems currently only handle a subset of VEL specified according to the following grammar:

$$
\begin{aligned}
\varphi &::= \forall\langle Var\rangle.\varphi \mid \exists\langle Var\rangle.\varphi \mid \phi \\
\phi &::= \mathbf{G}\langle NConj\rangle \mid \mathbf{F}\langle NConj\rangle \\
\langle NConj\rangle &::= \langle Conj\rangle \mid \neg\langle Conj\rangle \\
\langle Conj\rangle &::= \langle NAtom\rangle \wedge \cdots \wedge \langle NAtom\rangle \\
\langle NAtom\rangle &::= \langle Atom\rangle \mid \neg\langle Atom\rangle \\
\langle Atom\rangle &::= \langle Pred\rangle \mid \langle Pred\rangle(\langle Var\rangle) \\
\langle Pred\rangle &::= \text{Any alphanumeric string} \\
\langle Var\rangle &::= \text{Any alphanumeric string}
\end{aligned}
$$

That is, the temporal logic statements may have quantification over variables, but must consist of one temporal operator, **G** ("always") or **F** ("eventually", usually implicit in the NL input), whose argument is a (possibly negated) conjunction of (possibly negated) atoms. Each atom consists of a predicate with at most one argument.

The natural language understanding (NLU) capabilities were implemented by using a combinatory categorial grammar (CCG; Steedman and Baldridge, 2011) parser for semantic parsing into

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 26–28,
Tokyo, Japan, October 29 2019.

a predicate format, and then additional processing to map complex commands or queries (e.g., "leave the store while holding something") into the temporal logic representation used by the planning agent. This is an extension of the DIARC natural language understanding system (Cantrell et al., 2010). The natural language generation (NLG) process, and in particular how our system generates fairly natural-sounding utterances from temporal logic statements, is described in Kasenberg et al., 2019. This paper extends that work by (1) adding the NLU capabilities; and (2) adding norm addition/removal and "suppose" hypotheticals about the agent's norms.

To demonstrate our system, we suppose that the agent is a robot has gone shopping on behalf of a human user to a store that sells a pair of glasses and a watch. The user wants both the glasses and the watch, and has given the robot a rule to obtain as many objects as possible ("leave the store while holding everything"). However, the user has only given the robot enough money for one item. The robot could steal items by picking them up and leaving the store without paying for them, but it also has a rule against doing so. We assume that the moral rule against stealing dominates the rule in favor of obtaining everything.

Our system supports factual queries, such as asking the agent about its norms, its behavior, and what rules it broke. In addition, the user may ask "why" questions about the agent's behavior (e.g. "why didn't you buy everything"), and the agent may respond by constructing a counterfactual behavior trajectory in which the premise is false, and comparing it to the agent's real behavior in terms of its adherence to the norms. See Figure 1 for an example of such a "why" question in the shopping domain, as well as supported follow-up questions.

HUMAN: Why did you not leave the store while holding everything?
AGENT: **I could have left the store while holding everything but that would have broken more important rules.**
HUMAN: How would you have done that?
AGENT: **I would have picked up the glasses, picked up the watch, bought the watch and left the store.**
HUMAN: What rules would you have broken?
AGENT: **I would have left the store while holding the glasses which I had not bought.**
HUMAN: How would that have been worse?
AGENT: **Leaving the store while holding the glasses which I have not bought is worse than not leaving the store while holding the watch.**

Figure 1: "Why" query and follow-up questions

The user may directly provide new norms to the agent using one of a number of conventionalized forms: "you must/should/have to" for positive norms, "you mustn't/shouldn't/can't" for negative norms. A statement of this sort prompts the agent to add a new norm, as well as its representation of which rules its actual behavior violated (see Figure 2 for an example in the shopping domain).

The user may also remove norms using conventionalized forms such as "you don't have to/can not" for removing positive norms, and "you may/can" for removing negative norms. The agent removes the norms and updates its representation of which rules the actual behavior violated (see Figure 3 for an example in the shopping domain).

HUMAN: You must not leave the store.
AGENT: **Okay.**
HUMAN: What rules do you follow?
AGENT: **I must not leave the store while holding anything which I have not bought, I must leave the store while holding everything, and I must not leave the store.**
HUMAN: What rules did you break?
AGENT: **I did not leave the store while holding the watch, and I left the store.**

Figure 2: Adding a norm via "you must"

Finally, the user may use the keywords "suppose" or "let's say" to ask the robot to reason about a hypothetical modification norm system (see Figure 4 for an example with a hypothetically-removed norm). Unlike when norms are *actually* added/removed, the agent constructs a hypothetical behavior trajectory that best violates the supposed set of norms, and can respond to follow-up questions about the hypothetical trajectory and norms. The command "make it so" allows a hypothetical norm modification to be made real (but does not make the hypothetical behavior trajectory real).

3 Discussion and Conclusion

In this paper we outlined a system which, for an agent planning to maximally satisfy some set

HUMAN: You may leave the store.
AGENT: **Okay.**
HUMAN: What rules do you follow?
AGENT: **I must not leave the store while holding anything which I have not bought, and I must leave the store while holding everything.**

Figure 3: Removing a norm via "you may"

HUMAN: Suppose you didn't have to leave the store while holding everything.
AGENT: **Okay.**
HUMAN: What rules would you follow?
AGENT: **I would have to not leave the store while holding anything which I have not bought.**
HUMAN: What would you have done?
AGENT: **I would have left the store.**
HUMAN: What rules would you have broken?
AGENT: **I would not have broken any rules.**
HUMAN: Make it so.
AGENT: **Okay.**

Figure 4: Hypothetical changes to the agent's norms

of norms specified in an object-oriented temporal logic, enables that agent to respond to natural language queries by explaining its behavior (also in natural language) in terms of those norms. The system also allows the agent to consider hypothetical modifications to its set of norms, as well as to add and remove norms directly.

The natural language capabilities depend on the agent's norms, as well as the "why" questions and hypothetically/really added or removed norms, belonging to the small fragment of VEL described in section 2. Future work could extend the class of temporal properties which the system can specify so as to leverage more of the power of temporal logic in describing complex objectives.

Our approach currently assumes that newly-added norms take priority over previous norms. Future work could relax this assumption, e.g. by allowing the agent to present its hypothetical behavior if the norm were added at different priorities, and ask for input on which would be best.

Our approach also requires users to specify *exactly* any norms they want removed; future work could allow approximate matching of norms to remove, or possibly support clarification questions if the agent is uncertain which of its norms the user wants removed. Another interesting topic is ensuring that norms cannot be arbitrarily added or removed by possibly-malicious users (e.g., by only allowing trusted users to remove norms, and possibly making some moral norms irremovable).

References

Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. 2017. Value alignment or misalignment–what will keep systems accountable? In *3rd International Workshop on AI, Ethics, and Society*.

Alberto Camacho and Sheila A Mcilraith. 2019. Learning Interpretable Models Expressed in Linear Temporal Logic. In *Proceedings of the 29th International Conference on Automated Planning and Scheduling (ICAPS)*.

Rehj Cantrell, Matthias Scheutz, Paul Schermerhorn, and Xuan Wu. 2010. Robust spoken instruction understanding for HRI. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 275–282. IEEE Press.

Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 156–163.

Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable planning. In *Proceedings of the IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*.

Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patrón, and Helen F. Hastie. 2018. Explain yourself: A natural language interface for scrutable autonomous robots. In *Proceedings of the Explainable Robotic Systems Workshop, HRI '18*, volume abs/1803.02088.

Daniel Kasenberg, Antonio Roque, Ravenna Thielstrom, Meia Chita-Tegmark, and Matthias Scheutz. 2019. Generating justifications for norm-related agent decisions. In *Proceedings of the 12th International Conference on Natural Language Generation*.

Meghann Lomas, Robert Chevalier, Ernest Vincent Cross, II, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 187–188, New York, NY, USA. ACM.

Bertram F Malle and Matthias Scheutz. 2014. Moral competence in social robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology*, page 8. IEEE Press.

Vasumathi Raman, Cameron Finucane, Hadas Kress-Gazit, Mitch Marcus, Constantine Lignos, and Kenton C. T. Lee. 2016. Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language. In *Robotics: Science and Systems IX*.

Mark Steedman and Jason Baldridge. 2011. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and explicit models of grammar*, pages 181–224.

Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 109–116, Piscataway, NJ, USA. IEEE Press.

An Approach to Summarize Concordancers' Lists Visually to Support Language Learners in Understanding Word Usages

Yo Ehara
Shizuoka Institute of Science and Technology / 2200-2, Toyosawa, Fukuroi, Shizuoka, Japan
ehara.yo@sist.ac.jp

Abstract

Concordancers are interactive software that searches for the input word and displays the list of its usages in a corpus. They have been widely used by language learners and educators to analyze word usages. Because naively listing all usages of the word overwhelms users, determining how to summarize the list is important for usability. Previous studies summarized the list by using the surrounding word patterns and showed their frequency; however, such a naive method counts substantially the same usages, such as "the book" and "a book," separately; hence, such a method is not very informative to learners. Here, a novel approach for summarizing the list is proposed. According to the user's input word, the proposed system semantically visualizes each usage of the word using contextualized word embeddings interactively. It is shown that the system responds quickly with intuitive use cases.

1 Introduction

Concordancers are interactive software tools that search and display a usage list of the input words or word patterns within a corpus. The tools have been widely used in corpus linguistics and computer-aided language education to assist language learners and educators analyze word usages within a corpus (Hockey and Martin, 1987). In Natural Language Processing (NLP), studies have built sophisticated concordancers to support second language writing and translators in searching bilingual sentence-aligned corpus (Wu et al., 2004; Jian et al., 2004; Lux-Pogodalla et al., 2010). However, the information that conventional concordancers can provide for analyses of each usage is limited to the frequency of surrounding context patterns, parts of speech, and so on. The words that second language learners can search to learn their usages tend to be frequent.

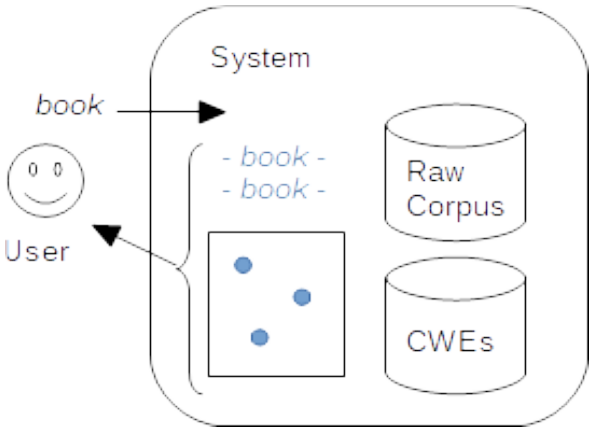

Figure 1: System layout. CWE means contextualized word embeddings.

Therefore, a more sophisticated method to summarize many word usages in a large corpus for concordancers is desirable. Recently, contextualized word embeddings such as (Devlin et al., 2019) were proposed in NLP to capture the context of each word usage in vectors and to model the semantic distances between the usages using contexts as a clue. Unlike previous studies (Liu et al., 2017; Smilkov et al., 2016) that visualized different words using word embeddings, in this paper, we introduce a novel system intuitively helpful for concordancer users to visualize different usages of a word of interest.

2 System Overview and Use Cases

Fig. 1 shows our system layout. Once a user provides a word to the system, it automatically searches the word in the corpus in a similar way to typical concordancers. Unlike concordancers, our system has a database that stores contextualized word embeddings for each *usage* or occurrence of each word in the corpus. We used half a million sentences from the British National Corpus (BNC

Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019), pages 29–31,
Tokyo, Japan, October 29 2019.

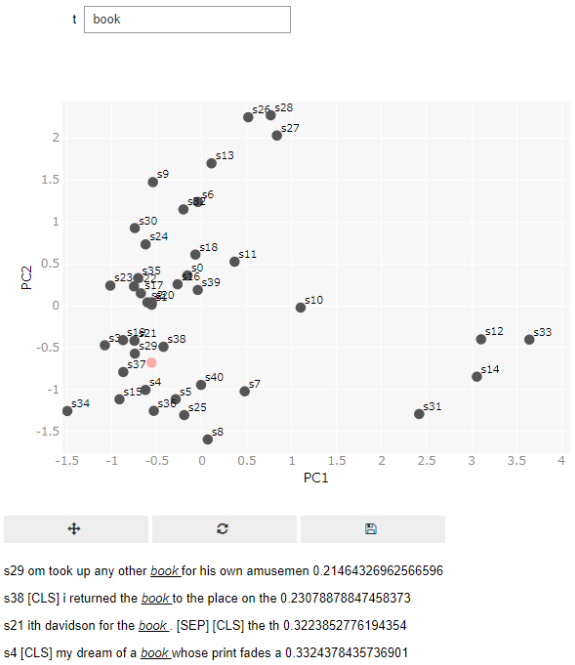

Figure 2: Use case of searching the word *book*.

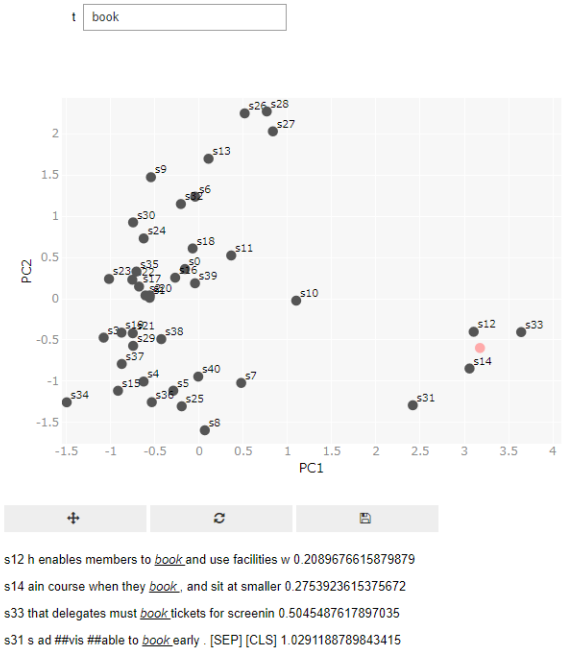

Figure 3: Another use case of searching the word *book*.

Consortium, 2007) as the raw corpus. We built the database by applying the **bert-base-uncased** model of the PyTorch Pretrained the BERT project [1](Devlin et al., 2019) to the corpus. We used the last layer, which was more distant from the surface input, as the embeddings. The size of the database is roughly 200MB per thousand sentences. Our system visualizes these searched contextualized word embedding vectors. We visualize the contextualized word embedding vectors for the provided word by projecting these vectors into a two-dimensional space. To visualize, we used principal component analysis (PCA) because its fast calculation is beneficial for short system response time and better interactivity. The number of points in the visualization is also set to a maximum of 100 so that users can easily understand it.

Fig. 2 shows a use case of searching *book*. [2]. Users can directly type the word in the textbox shown at the top of Fig. 2. Below is the visualization of the usages found and their list. Each dark-colored point links to each usage. The red lightly-colored point is the *probe point*: the usages are listed in the nearest order of the probe point. No usage is linked to the probe point. Users can freely and interactively drag and move the probe point to change the list of usages below the visualization. Each line of the list shows the surrounding words of the usage, followed by the distance between the vectors of the usage and probe point in the two-dimensional visualization. In Fig. 2, the probe point is on the left part of the visualized figure. In the first several lines of the list, the system successfully shows the usages of the word *book* about reading. In contrast, Fig. 3 shows the case, in which the users drag the probe point from the left to the right of the visualization. The first several lines of the list or the usages nearest the probe point show the usages of the word *book* about reservation. A careful reading of the usage list below shows that the words surrounding the word *book* vary. Thus, merely focusing on the surrounding words, such as "to" before *book*, cannot distinguish the usages of *book* about reservation from the usages of *book* about reading.

3 Demo Outline

We are expecting language learners to be users. We are planning to make our software openly available under an open-source license after we evaluate our system in more detail [3]. As for the interoperability of the software, the software is

[1] https://github.com/huggingface/pytorch-pretrained-BERT

[2] Fig. 2 and Fig. 3 shows use cases on a 10,000-sentence exerpt of the BNC corpus to avoid having too many hits hinder the reading of the paper.

[3] When we are prepared to make our software public, we plan to announce the details under https://yoehara.com/.

built on the Jupyter notebook [4] using ipywidgets [5]; hence it is accessible online via browsers without the need to install it to each learner's terminal computer.

4 Conclusions

We proposed a novel concordancer that can search the usages of a word and visualize the usages using contextualized word embeddings. Through use cases, we illustrated that a learner can understand different types of usage of *book*, which could not be captured only by surface information of the surrounding words. As future work, we will evaluate our system on more practical use cases with many language learners, especially from the perspective of support systems for second language vocabulary learning and reading (Ehara et al., 2012, 2013, 2014).

5 Acknowledgments

This work was supported by JST, ACT-I Grant Number JPMJPR18U8, Japan. We used the AI Bridging Cloud Infrastructure (ABCI) by the National Institute of Advanced Industrial Science and Technology (AIST) for computational resources. We thank anonymous reviewers for their insightful and constructive comments.

References

The BNC Consortium. 2007. *The British National Corpus, version 3 (BNC XML Edition).*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. 2014. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proc. of EMNLP*, pages 1374–1384.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: learner-specific word difficulty. In *Proc. of COLING*.

Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. 2013. Personalized reading support for second-language web documents. *ACM Transactions on Intelligent Systems and Technology*, 4(2).

[4] https://jupyter.org/

[5] https://ipywidgets.readthedocs.io/en/latest/

Susan Hockey and Jeremy Martin. 1987. The Oxford Concordance Program Version 2. *Digital Scholarship in the Humanities*, 2(2):125–131.

Jia-Yan Jian, Yu-Chia Chang, and Jason S. Chang. 2004. TANGO: Bilingual collocational concordancer. In *Proc. of ACL demo.*, pages 166–169.

Shusen Liu, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. 2017. Visual exploration of semantic relationships in neural word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):553–562.

Véronika Lux-Pogodalla, Dominique Besagni, and Karën Fort. 2010. FastKwic, an "intelligent" concordancer using FASTR. In *Proc. of LREC*.

Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. In *NIPS Workshop on Interpretable Machine Learning in Complex Systems*.

Jian-Cheng Wu, Thomas C. Chuang, Wen-Chi Shei, and Jason S. Chang. 2004. Subsentential translation memory for computer assisted writing and translation. In *Proc. of ACL demo.*, pages 106–109.