NL4XAI 2020

2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence

Proceedings of NL4XAI

18 December 2020
Dublin, Ireland

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.

Interactive *Natural Language* Technology for eXplainable Artificial Intelligence

NL4XAI

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 978-1-952148-56-9

Introduction

Welcome to the Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020)!

This workshop takes place co-located with the International Conference on Natural Language Generation (INLG2020), which is supported by the Special Interest Group on NLG of the Association for Computational Linguistics. INLG 2020 was due to be held in Dublin (Ireland), 15 December - 18 December, 2020. However, due to covid-19 INLG2020 became a fully online event. The NL4XAI workshop is scheduled by December 18. This is the second of a series of workshops to be organized in the context of the European project NL4XAI (https://nl4xai.eu/).

NL4XAI is the first European Training Network on Natural Language (NL) and Explainable Artificial Intelligence (XAI). This network is funded by the Horizon 2020 research and innovation programme, through a Marie Skłodowska-Curie grant, in the framework of the European Union's bet for XAI. NL4XAI is a joint academic-industry research network, that brings together 18 beneficiaries and partners from 6 different countries (France, Malta, Netherlands, Poland, Spain, and United Kingdom). They work together with the aim of making AI self-explaining and contributing to translate knowledge into products and services for economic and social benefit. The goal is to produce intelligent machines able to explain their behavior and decisions through interactive explanations in NL, just as humans naturally do. NL technologies, both NL Generation (NLG) and NL Processing (NLP) techniques, are expected to enhance knowledge extraction and representation for XAI through human-machine interaction (HMI). Eleven Early Stage Researchers (ESRs) face different but complementary research challenges to accomplish this goal. The NL4XAI network offers a unique research environment providing ESRs with an excellent structured training programme.

This workshop provides attendants with a forum for: (1) disseminating and discussing recent advances on XAI; (2) identifying challenges and exploring potential transfer opportunities between related fields; (3) generating synergy and symbiotic collaborations in the context of XAI, HMI and NL technologies.

We received 17 submissions (16 regular papers and 1 demo). Twelve regular submissions were accepted to be included in the program after a double blind peer review. In addition, NL4XAI 2020 includes two outstanding invited speakers. The first invited speaker, in the morning, will be Prof. Dr. Emiel Khramer (*Tilburg center for Cognition and Communication (TiCC)*). He will talk about explaining health information automatically. The second invited speaker, in the afternoon, will be Dr. Eirini Ntoutsi (*Leibniz Universität Hannover & L3S Research Center*). She will talk about bias in AI-systems. In addition, the program includes a round table regarding open research challenges. We are glad to have Emiel Khramer, Eirini Ntoutsi and Albert Gatt as panelists.

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

Jose M. Alonso and Alejandro Catala
NL4XAI 2020 Organizers

Workshop Organizers:

Jose M. Alonso, CiTIUS, University of Santiago de Compostela
Alejandro Catala, CiTIUS, University of Santiago de Compostela

Program Committee:

Jose M. Alonso, CiTIUS, University of Santiago de Compostela
Katarzyna Budzynska, Warsaw University of Technology
Alberto Bugarin, CiTIUS, University of Santiago de Compostela
Alejandro Catala, CiTIUS, University of Santiago de Compostela
Kees van Deemter, Utrecht University
Pablo Gamallo, CiTIUS, University of Santiago de Compostela
Claire Gardent, CNRS/LORIA
Albert Gatt, University of Malta
Marcin Koszowy, Warsaw University of Technology
Jordi Levy, IIIA - CSIC
Chenghua Lin, University of Sheffield
Simon Mille, Universitat Pompeu Fabra
Nir Oren, University of Aberdeen
Martin Pereira-Fariña, Dept. de Filosofia e Antropoloxia, University of Santiago de Compostela
Alejandro Ramos-Soto, University of Santiago de Compostela
Ehud Reiter, University of Aberdeen, Arria NLG plc.
Carles Sierra, IIIA - CSIC
Mariët Theune, Human Media Interaction, University of Twente
Nava Tintarev, Technische University of Delft

Invited Speakers:

Emiel Krahmer, Tilburg center for Cognition and Communication (TiCC)
Eirini Ntoutsi, Leibniz Universität Hannover & L3S Research Center

Panelists:

Emiel Krahmer, Tilburg center for Cognition and Communication (TiCC)
Eirini Ntoutsi, Leibniz Universität Hannover & L3S Research Center
Albert Gatt, Institute of Linguistics and Language Technology, University of Malta (UM)

Table of Contents

Workshop Program

Friday, December 18, 2020 (GMT)

11:00–11:05 Welcome

11:05–12:00 Keynote Talk (Prof. Dr. Emiel Krahmer)

12:00–12:30 Oral Presentations I

12:30–12:50 Virtual Coffee/Tea Break

12:50–14:20 Oral Presentations II

14:20–15:00 Lunch break

15:00–16:00 Keynote Talk (Dr. Eirini Ntoutsi)

16:00–16:30 Oral Presentations III

16:30–16:50 Virtual Coffee/Tea break

16:50–17:20 Oral Presentations IV

17:20–18:10 Round table: "XAI Open Challenges"

18:10–18:20 Concluding remarks and farewell

Invited talk

Automatically explaining health information

Emiel Krahmer
e.j.krahmer@tilburguniversity.edu

Tilburg center for Cognition and Communication (TiCC)

Abstract

Modern AI systems automatically learn from data using sophisticated statistical models. Explaining how these systems work and how they make their predictions therefore increasingly involves producing descriptions of how different probabilities are weighted and which uncertainties underlie these numbers. But what is the best way to (automatically) present such probabilistic explanations, do people actually understand them, and what is the potential impact of such information on people's wellbeing?

In this talk, I adress these questions in the context of systems that automatically generate personalised health information. The emergence of large national health registeries, such as the Dutch cancer registry, now make it possible to automatically generate descriptions of treatment options for new cancer patients based on data of comparable patients, including health and quality of life predictions following different treatments. I describe a series of studies, in which our team has investigated to what extent this information is currently provided to people, and under which conditions people actually want to have access to these kind of data-driven explanations. Additionally, we have studied whether there are different profiles in information needs, and what the best way is to provide probabilistic information and the associated undertainties to people.

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 1–2,
Dublin, Ireland, 18 December 2020.

Invited talk

Bias in AI-systems: A multi-step approach

Eirini Ntoutsi
ntoutsi@kbs.uni-hannover.de

Leibniz Universität Hannover & L3S Research Center

Abstract

Algorithmic-based decision making powered via AI and (big) data has already penetrated into almost all spheres of human life, from content recommendation and healthcare to predictive policing and autonomous driving, deeply affecting everyone, anywhere, anytime. While technology allows previously unthinkable optimizations in the automation of expensive human decision making, the risks that the technology can pose are also high, leading to an ever increasing public concern about the impact of the technology in our lives. The area of responsible AI has recently emerged in an attempt to put humans at the center of AI-based systems by considering aspects, such as fairness, reliability and privacy of decision-making systems.

In this talk, we will focus on the fairness aspect. We will start with understanding the many sources of bias and how biases can enter at each step of the learning process and even get propagated/amplified from previous steps. We will continue with methods for mitigating bias which typically focus on some step of the pipeline (data, algorithms or results) and why it is important to target bias in each step and collectively, in the whole (machine) learning pipeline. We will conclude this talk by discussing accountability issues in connection to bias and in particular, proactive consideration via bias-aware data collection, processing and algorithmic selection and retroactive consideration via explanations.

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 3–4,
Dublin, Ireland, 18 December 2020.

Content Selection for Explanation Requests in Customer-Care Domain

Luca Anselma[♡] **Mirko Di Lascio**[♡] **Dario Mana**[♣]
Alessandro Mazzei[♡] **Manuela Sanguinetti**[♡♢]
[♡]Dipartimento di Informatica, Università degli Studi di Torino, Italy [♣]TIM, Torino, Italy
[♢]Dipartimento di Matematica e Informatica, Università degli Studi di Cagliari, Italy
[♡]{first.last}@unito.it, [♢]{first.last}@unica.it
[♣]{first.last}@telecomitalia.it

Abstract

This paper describes a content selection module for the generation of explanations in a dialogue system designed for customer care domain. First we describe the construction of a corpus of dialogues containing explanation requests from customers to a virtual agent of a telco, and second we study and formalize the importance of a specific information content for the generated message. In particular, we adapt the notions of *importance* and *relevance* (Biran and McKeown, 2017) in the case of schematic knowledge bases.

1 Introduction

Customer care is one of the application domains where Dialogue Systems (DSs) represent an emerging technology used by many big companies to satisfy customer requests (MITTR, 2018). Customer care dialogues can have a specific linguistic characterization (Oraby et al., 2019), and often the customer preferences lean toward short dialogues (Demberg et al., 2011). Moreover, in the customer care domain the users' requests often regard some form of *explanation* about their past transactions with the company. To provide explanations, commercial DSs often provide long lists of data entries extracted from databases containing company-customer relationship data. Therefore, there is the necessity to give some form of *priority* to data entries to present just – or to give more prominence to – the information that is most relevant for the user (Demberg et al., 2011).

Most commercial DSs follow the classical cascade architecture $NLUnderstanding \leftrightarrow DialogueManager \leftrightarrow NLGeneration$ (McTear et al., 2016). This architecture relies, as a working hypothesis, on the assumption that most of necessary information is provided by the user utterance. However, this assumption is sometimes false or only partially true. For instance, in the sentence "*Scusami ma vorrei sapere come mai mi vengono fatti certi addebiti?*" ("Excuse me, I'd like to know why I'm charged certain fees?"), even a very advanced NLU module can produce only vague information about the user's request to the dialogue manager. Indeed, to provide an appropriate response, the dialogue manager might need to ask for additional clarification or, in alternative, to access some contextual information to obviate the lack of linguistic information. In the case of customer care, this contextual information can be found as schematic knowledge bases arising from databases. As a result, when linguistic information is scarce (or *absent* in the case of ungrammatical/incomprehensible input) retrieving and giving priority to contextual information in DSs is essentially a problem of *content selection* (Reiter and Dale, 2000). Therefore, as a working hypothesis, in this paper we consider negligible the linguistic input given by the user. However, also when the linguistic input is comprehensible, a good balance between the information carried by the linguistic input and by the specific domain context is a key goal for the dialogue manager.

The idea to use NLG techniques for explaining rationales inside data is a topic that is drawing growing attention (Reiter, 2019). One of the few papers providing a quantitative evaluation of explanations was produced by Biran and McKeown (2017). In this work the authors proposed a model for quantifying the relevance of a feature for a specific class of machine learning algorithms, i.e. linear classifiers. The authors introduced two notions, *importance* and *effect*, to evaluate the relevance of a feature in the general classification model and for a specific classification instance respectively. The basic idea was to determine the narrative role of a feature based on the combination of its importance and its effect; for example, a feature may

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 5–10,
Dublin, Ireland, 18 December 2020.

have the narrative role of *exceptional evidence* in the case of low importance and high effect. In this way, the authors have been able to communicate the key data elements into core messages for an explanation (*justification* in their terminology).

In this paper, we present some initial results of an ongoing study on the design of a generation module of a DS in the domain of telco customer care. We focus our study on customers' requests of *explanations* (Reiter, 2019). The study presented here, in fact, is part of a wider project that aims to improve the answers provided by a virtual agent of an online customer service, by creating a NLG system that could also take into account various dimensions in the generation process, such as possible errors in the conversations (see e.g. Bernsen et al. (1996), Martinovsky and Traum (2003), Higashinaka et al. (2015)) and the presence of emotions (especially negative ones) in the user messages. At this stage of the project, we use the model presented in Biran and McKeown (2017) to give relevance to the *content units* in the knowledge about the customer. In particular, we adapt the definition of the narrative roles for *importance* and *effect* to the case of a knowledge base consisting of database entries.

This paper provides two main specific contributions: (1) the analysis of a corpus consisting of real dialogues containing explanation requests (Section 2), (2) the proposal of a content-selection procedure based on narrative roles in explanation when the DS contextual data is a schematic knowledge base arising from a database (Section 3). In the final Section of the paper we discuss these contributions in relation to our ongoing work.

2 Building a corpus of explanation requests

This study builds upon the analysis of a corpus of dialogues between customers and a virtual agent for customer service developed by an Italian telecommunications company (Sanguinetti et al., 2020). The dialogues, which take place by means of a textual chat, mainly deal with requests for commercial assistance, both on landline and mobile phones. For the purpose of this study, the corpus created was extracted by selecting, from a sample of dialogues held over 24 hours, a reduced subset that included requests for explanations from customers. The selection criteria were conceived so as to include all the dialogues where at least one message from the user contained a clearly stated

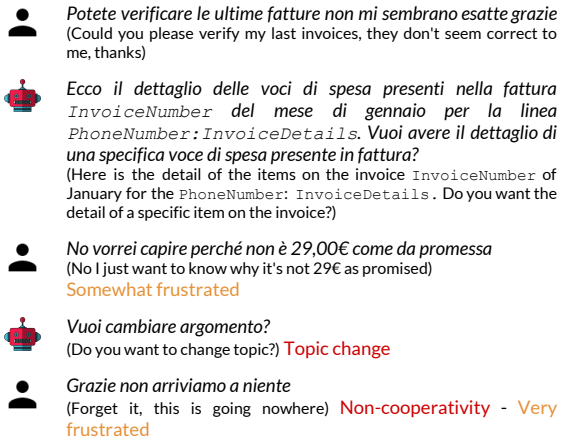

Figure 1: An annotated dialogue with additional annotation layers: errors (red) and emotions (orange).

request for explanation. A simple string search method was thus carried out to filter such kind of dialogues, using the following strings: *sapere/capire perché*[1] ("know/understand why") and *come mai* ("why/how come"). The resulting corpus consists of 142 dialogues, with an average of 11 turns per dialogue, and an average length of 9 tokens in customer messages and 38 tokens in the bot messages. Such difference in the message length is due to the way the assistant's responses are currently structured, in that they usually include detailed information e.g. on invoice items or options available, while, on the other hand, customer's messages are most often quite concise. Also, the relatively high number of turns per dialogue might be explained with the high occurrence in the corpus of repeated or rephrased messages, both by the virtual agent and the user, due to recurring misunderstandings on both sides. The corpus underwent an annotation process that involved multiple, complementary, dimensions, such as errors in conversation and emotions (see Figure 1 for an example).

The explanation request and its sub-types have been included as one of such dimensions and we mainly focused our attention on these in this phase of the study. The types of requests for explanations in this collection reflect the different kinds of problems typically encountered with a telephone operator. Based on a preliminary analysis of the corpus, we distinguished 5 main types of requests plus a generic category that includes a variety of cases that is more heterogeneous and not classifiable according to the main types. Hence, we identified requests for explanations or clarifications

[1] Variants as *perchè, xk, xkè* have been used too.

regarding the following topics: (1) charges in the invoice or in the phone credit (about 52% of cases), (2) timing and methods of receipt of the invoice (10.5%), (3) unpaid invoice reminders erroneously received (10.5%), (4) currently active promotions (8%), (5) payment methods (5%). The remaining cases (14%) were included in the more generic "Other" class. Starting from this analysis we thus defined a reduced set of possible *scenarios*, i.e. prototypical situations that can be found in the dialogues and grouped together according to similar characteristics. For illustrative purposes, we describe in this paper the three main scenarios defined for the first request type, i.e the one regarding undue or unclear charges, being by far the most frequent case of request. In Scenario 1 (31% of the occurrences) the customer asks for an explanation about a higher charge with respect to previous ones, also providing specific information on the amount charged; in Scenario 2 (58% of the occurrences), a charge in the account is claimed, but no further information is provided in the user's message. In Scenario 3 (11% of the occurrences) the customer asks for an explanation about a negative balance.

3 Importance and Effect for Content Selection in the Customer Care Domain

We consider three scenarios arising form the corpus analysis (Section 2). Formally, each scenario consists of a set of sequences of transactions, where a transaction is a money transfer operation between a customer and the company (i.e., an amount paid for a certain service). As a result, each transaction sequence represents the different amounts paid along a time period for a specific service (transaction type). To determine the importance and the effect of a transaction sequence, we assume to know all the transactions on the user's account in the last seven months.

It is worth pointing out that the two most important elements in this specific context are money and time. Therefore, we want to formalize the intuition that the *importance* (in Biran and McKeown's terminology) of a telco service can be associated with the amount of money that the user usually spends for such service, while its *effect* can be associated with the amount of money that the user spent for the service in the last month.

We thus define the *importance of a transaction sequence* as the mean of the normalized values of the transactions in the past six months. Moreover, we define the *effect of a transaction sequence* as the normalized value of the transactions in the current month. Normalization is carried out by dividing the amount of the transactions by the maximum amount that the user has paid for that transaction. An important point in the Biran and McKeown (2017) model is the procedure for transforming importance/effect numeric values in the discrete {low, high} values. In accordance to the original model, we determine the smallest subset H of transaction sequences such that the sum of their importance/effect values is greater than the 75% of the total importance/effect. When such a smallest subset is not unique, we consider the union of all the smallest subsets. Note that the value of 75% has been set in order to adhere to the original model, that has been proposed in (Biran and McKeown, 2017) without a specific motivation. However, we consider this limit as a tunable value that should be empirically validated on the specific domain.

In the following discussions we analyse the scenarios for three common requests of explanation. We separately analyse these scenarios but not that they are not mutually exclusive. It is worth noting that a complete NLG architecture could account their possible coexistence by using some form of syntactic or semantic aggregation.

Finally, note that in the discussion on these scenarios we are completely neglecting both the linguistic information arising from the dialogue (the user's question) as well as any kind of information on the customer. In other words, we are inferring the customer's explanation request as a content selection task only, without taking user utterances and user model into account. As a matter of fact, there are some cases where the user searches for a complete information about its transactions: for instance, the user wants to review all the transactions of the last months. In this case, the linguistic input should trigger the dialogue manager and the NLG system to provide information on transactions with *normal* evidence after the information on exceptional evidence. In contrast, there are situations such that the user wants to have only a short summary on its transactions and in this case the NLG system should only provide information on transactions with exceptional evidence. In future research, we plan to study how to merge the linguistic, the domain and the user model information.

	M1	M2	M3	M4	M5	M6	M7
S1	10	10	10	10	10	10	10
S2	0	0	0	0	0	2	2

Table 1: The distribution of transactions along the current month (M7) and the previous six months (M1-M6) for Scenario 1.

	M1	M2	M3	M4	M5	M6	M7
S1	9.99	9.99	9.99	9.99	9.99	9.99	9.99
S2	0	0	0	0	2	2	2, 2
S3	0	0	0	0	0	0	1.59

Table 2: The distribution of transactions for Scenario 2.

3.1 Scenario 1

Scenario 1 represents a typical situation of a user requesting for an explanation about a total charge in the current month higher than the ones in the previous months. The interaction between the DS and the user starts with a short message: *Salve vorrei sapere perché mi sono stati presi* 12 € *invece che dieci dall'ultima ricarica (Hi I'd like to know why you got* 12 € *instead of ten since last top-up).*

We assume for this scenario that the user paid for two services[2] (that are transaction sequences, see Table 1). In particular, a transaction of 10€ is present in each month (M1-M7) for S1, while a transaction of 2€ is present only in months $M6$ and $M7$ for S2. By using the data in Table 1, we can calculate the importance and the effect for S1 and S2. The importance of S1 is $(10/10 + 10/10 + 10/10 + 10/10 + 10/10 + 10/10)/6 = 1$, while the importance of S2 is $(0/2 + 0/2 + 0/2 + 0/2 + 0/2 + 2/2)/6 = 0.17$, thus the sum of the importance values is 1.17 and its 75% value is 0.88. The smallest subset H_I such that the sum of the importance values of the transactions is greater than 0.88 is $H_I = \{S1\}$. As a result, S1 has high importance, while S2 has low importance. The effects of S1 and S2 are $10/10 = 1$ and $2/2 = 1$, therefore the sum of the effect values is 2 and its 75% is 1.5. The smallest subset H_E such that the sum of the effect values is greater than 1.5 is $H_E = \{S1, S2\}$, hence S1 and S2 have both high effect. Thus, combining the discrete values of importance and effect, S1 is a normal evidence since it has high importance and high effect, and S2 is an exceptional evidence since it has low importance and high effect. This exceptional evidence captures the intuition that S2 is more informative than S1 in Table 1. As a consequence, S2 will have a central role in the requested explanation.

[2]Note that the trivial solution to return both contents does not solve the problem of assigning them a priority in presentation.

3.2 Scenario 2

Scenario 2 represents a user requesting an explanation about some specific charges (*Scusami ma vorrei sapere come mai mi vengono fatti alcuni addebiti — Sorry but I'd like to know why there are some charges*).

This scenario has three transaction sequences: S1, with an amount of 9.99€ (M1-M7), S2 with an amount of 2€ (M5-M7, appearing twice in M7), and S3 with an amount of 1.59€ (M7) (see Table 2). From this data, we calculate importance and effect for S1, S2 and S3, and their narrative roles as described previously. The importance of S1 is 1, the importance of S2 is 0.33 and the importance of S3 is 0. The sum of the importance values is 1.33 and its 75% is 0.99. The smallest subset H_I such that the sum of the importance values is greater than 0.99 is $H_I = \{S1\}$, so S1 has high importance, while S2 and S3 have low importance. The effect of a transaction sequence is given by the values in the current month: S1 and S3 effect is 1 and S2 effect is 2. The sum of the effect values is 4 and its 75% is 3. The smallest subset H_E such that the sum of the effect is greater than 3 is $H_E = \{S1, S2, S3\}$, hence S1, S2 and S3 have all high effects. As a result, combining the discrete values of importance and effect S1 is a normal evidence and S2 and S3 are both exceptional evidences.

3.3 Scenario 3

Scenario 3 represents a user requesting an explanation about a negative balance (*Buongiorno,vorrei sapere perché ho il credito in negativo, nonostante abbia fatto una ricarica da* 15€ *proprio stamattina — Good morning, I'd like to know why I have a negative balance, despite I made a* 15€ *recharge just this morning*).

This user has three transactions sequences: S1 with an amount of 13€ (M1-M3) and 15€ (M4-M7), S2 with an amount of 0.9€ (four times in M7), and S3 with an amount of 1.99€ (M7) (see Table 3). From these data, we can calculate importance and effect for S1, S2 and S3 and their narrative roles

	M1	M2	M3	M4	M5	M6	M7
S1	13	13	13	15	15	15	15
S2	0	0	0	0	0	0	0.9, 0.9, 0.9, 0.9
S3	0	0	0	0	0	0	1.99

Table 3: The distribution of transactions for Scenario 3.

as previously described. The importance of S1 is 0.94, the importance of S2 and S3 is 0. The sum of the importance values is 0.94 and its 75% value is 0.71. The smallest subset H_I such that the sum of the importance values is greater than 0.71 is $H_I = \{S1\}$, so S1 has high importance, while S2 and S3 have low importance. S1 and S3 effect is 1, while S2 effect is 4. The sum of the effect values is 6 and its 75% value is 4.5. The smallest subset H_E such that the sum of the effect values of the transaction sequences in the subset is greater than 4.5 can be $\{S1, S2\}$ or $\{S2, S3\}$. The subset H_E is the union the two cases, i.e. $H_E = \{S1, S2, S3\}$, hence S1, S2 and S3 have high effect. Thus, S1 is a normal evidence, and S2 and S3 are exceptional evidences.

4 Conclusions and Future Work

This paper reports the first results of an ongoing study on the role of NLG for a DS in the customer care domain. We provided a corpus analysis that shed some light on the customer requests regarding explanations[3]. Moreover, we adapted the model proposed in Biran and McKeown (2017) for narrative roles in explanation for this specific kind of input data. In this way, we designed a content selection procedure accounting for *evidence* of data.

We are working on the inclusion of the content selection procedure described in this paper into a complete NLG architecture for DS. In this linguistically sound NLG architecture, we use a simple rule-based sentence planner (Anselma and Mazzei, 2018) in combination with the Italian version of SimpleNLG (Mazzei et al., 2016) for generating messages that give emphasis and priority to the content elements with high evidence. For instance, in this architecture we can decide to generate final messages that contain only (or mention primarily) contents with exceptional evidence.

As a future work, we are designing a user-based comparative evaluation of the DS exploiting the complete NLG architecture following the schema adopted in (Demberg et al., 2011). The idea is to show both a real dialogue from the corpus and a dialogue obtained with the complete NLG architecture, and to ask users to rate each dialogue and compare them by using a number of Likert-scale questions.

[3]We are currently working on the anonymization of the corpus in order to publicly release it.

Acknowledgment

This research has been partially funded by TIM s.p.a. (Studi e Ricerche su Sistemi Conversazionali Intelligenti, CENF CT RIC 19 01).

References

Luca Anselma and Alessandro Mazzei. 2018. Designing and testing the messages produced by a virtual dietitian. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 244–253.

Niels Ole Bernsen, Laila Dybkjær, and Hans Dybkjær. 1996. User errors in spoken human-machine dialogue. In *Proceedings of the ECAI '96 Workshop on Dialogue Processing in Spoken Language Systems*.

Or Biran and Kathleen McKeown. 2017. Human-centric justification of machine learning predictions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1461–1467.

Vera Demberg, Andi Winterboer, and Johanna D. Moore. 2011. A strategy for information presentation in spoken dialog systems. *Computational Linguistics*, 37(3):489–539.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015. Fatal or not? finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2248, Lisbon, Portugal. Association for Computational Linguistics.

Bilyana Martinovsky and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *Proceedings of the ISCA Workshop on Error Handling in Dialogue Systems*.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192, Edinburgh, UK. Association for Computational Linguistics.

Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface: Talking to Smart Devices*, 1st edition. Springer Publishing Company, Incorporated.

MIT Technology Review Insights MITTR. 2018. Humans + bots: Tension and opportunity.

Shereen Oraby, Mansurul Bhuiyan, Pritam Gundecha, Jalal Mahmud, and Rama Akkiraju. 2019. Modeling and computational characterization of twitter customer service conversations. *ACM Trans. Interact. Intell. Syst.*, 9(2–3).

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

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

Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Marco Scalerandi, Dario Mana, and Rossana Simeoni. 2020. Annotating errors and emotions in human-chatbot interactions in Italian. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 148–159, Barcelona, Spain. Association for Computational Linguistics.

ExTRA: Explainable Therapy-Related Annotations

Mat Rawsthorne
University of Nottingham

Tahseen Jilani
HDR-UK
Digital Research Service
University of Nottingham

Jacob Andrews
University of Nottingham

Yunfei Long
University of Essex

Jérémie Clos
University of Nottingham

Sam Malins
Institute of Mental Health
University of Nottingham

Daniel Hunt
University of Nottingham

Abstract

In this paper we report progress on a novel explainable artificial intelligence (XAI) initiative applying Natural Language Processing (NLP) with elements of co-design to develop a text classifier for application in psychotherapy training and practice. The task is to produce a tool that will automatically label psychotherapy transcript text with levels of interaction for patient activation in known psychological processes. The purpose is to enable therapists to review the effectiveness of their therapy session content. We use XAI to increase trust in the model's suggestions and predictions of the client's outcome trajectory. After pre-processing of the language features extracted from professionally annotated therapy session transcripts, we apply a supervised machine learning approach (CHAID) to classify interaction labels (negative, neutral or positive in terms of patient activation). Weighted samples are used to overcome class imbalanced data. The results show this initial model can make useful distinctions among the three labels of patient activation with 74% accuracy and provide insight into its reasoning. This ongoing project will additionally evaluate which XAI approaches are best for increasing the transparency of the tool to end users and explore whether direct involvement of stakeholders improves usability of the XAI interface and therefore trust in the solution.

1 Introduction

It takes a lot of manual effort to quality-assure psychotherapy sessions (Tseng et al., 2017), and therefore assessments of quality are rarely used routinely in psychotherapy practice. This work seeks to produce a tool that can automatically code psychotherapy transcripts, in line with a coding scheme developed by psychotherapists, known to characterise predictors of recovery (Malins et al., 2020a). The tool is also being developed to present explanations of the reasons for the coding decisions it makes. Explaining algorithms to those taking actions based on their outputs is recognised as good practice in data-driven health and care technology (DHSC, 2019). The ExTRA-PPOLATE [1] project is the first step in building tools to optimise scarce resources for provision of mental healthcare (Lorenzo-Luaces et al., 2017) by enabling therapists to adhere to good practice (Waller and Turner, 2016) and deliver care tailored to the patient (Delgadillo et al., 2016).

2 Overall Aims of Programme

The long-term objectives that we aim to achieve throughout our programme are threefold:

Aim 1 To build the foundation for unobtrusive, objective, transdiagnostic measures of patient activation.

Aim 2 To understand the practical trade-offs between classifier accuracy and explainability.

Aim 3 To explore the relationship amongst co-production, transparency and trust in algorithm-informed clinical decision making.

3 Methods

Core project team members were separately surveyed as to their initial hypotheses for key language markers of client-therapist interaction that is deemed helpful, focusing on generating features from different perspectives (see Table 1). These were then reviewed by the whole team and coded into a Python script to extract them from a corpus of transcripts of 120 health anxiety sessions. This created a simple model for identifying key interaction-types of interest (engagement in particular types of conversation) which are predictive of

[1] Explainable Therapy Related Annotations: Patient & Practitioner Oriented Learning Assisting Trust & Engagement

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NLXAI 2020), pages 11–15,
Dublin, Ireland, 18 December 2020.

clinical outcomes (Malins et al., 2020a) and could be compared to detailed labels that had been applied to the data by specialist raters in nVivo using the Clinical Interaction Coding Scheme (CICS) (Malins et al., 2020b). Further information on this coding scheme is provided in Appendix A.

4 Data Analysis

Data distribution and model selection The data was skewed, and it was necessary to collapse some similar categories to ensure sufficient representation. We employed Chi-square Automatic Interaction Detection (CHAID), a type of decision tree (DT) classification model that can handle both categorical and numeric data sets. It does not require common statistical assumptions such as normality and non-collinearity (Kass, 1980). For imbalanced data, DT models allow weighting samples according to their importance. A sub-category of the outcome variable having smaller number of samples is assigned higher weight than as compare to other category with larger number of samples. Since positive and neutral category ratings were more common in the dataset than negative ratings, negatively categorised data were weighted for balance.

How Decision Trees Work DT models work by recursively partitioning the samples into a number of subsets. The starting node (at the top of the tree) is termed as "root". Any node with outgoing nodes is termed as an internal node, while the nodes without further branches are called "leaves". At each node, the Chi-square test for association is applied and the variable having the strongest association with the outcome variable is selected for further split into leaves. The chosen variable is the one that expresses the strongest discrimination between the different levels of outcome variable. The algorithm keeps dividing the full data set into subsets using the depth-first approach until the stopping criterion is not met (Magidson, 1994).

Validation For internal validation of the model or when no validation data set is available, the model can perform K-fold cross validation. Finally, the results from different K folds were merged to produce a single DT estimation. DT models also offer tree pruning and feature selection based on the Chi-Squared test to prevent overfitting of the model. A "minimum cases" criteria is used for deciding further split of a branch. Discrimination of the original and cross-validated models was evaluated through the generation of Receiver Operating Characteristic (ROC) curves and calculation of C-statistics.

5 Initial Findings

CHAID label classification results are summarised in the classification matrix in Table 2. The overall accuracy of the model was 74% with the highest correct sample classified in the Neutral category. There were a total of 681 negative labels out of a total of 25,823 samples (2.6%). Of these, 60.4% were correctly classified. A larger total of 16,713 samples were recorded for positive labels (64.7% of the total), with a correct classification rate of 69.5%. The performance of the classification could be further enhanced through a more detailed exploration of the language features from the session transcripts, using improved oversampling techniques such as SMOTE and deeper machine learning modelling such as random forest and convolutional neural networks. Furthermore, the interdisciplinary engagement with the data has already helped deepen understanding of both the CICS framework and the classifier model (Páez, 2019) and generated ideas for their refinement.

6 Tool Development

The project uses a fusion of techniques to apply Responsible Research & Innovation (RRI) to the tool's development, specifically:

Incorporating a range of perspectives at multiple levels: The core project team combines the lived experience of a Service User Researcher and Involvement Volunteers (skilled in instrumentation design and plain English summaries) from the Institute of Mental Health, with specialist Clinical Psychology knowledge, Statistical Machine Learning, Psychometrics, Computer Science and Corpus Linguistics expertise. This diversity of experts in the formal and informal language of mental health provide triangulation to ensure the methods and findings make sense (Ernala et al., 2019). Additionally we engaged a Patient & Practitioner Reference Group (PPRG), comprised of 12 people, balanced across key stakeholder groups: patients and carers, clinical psychologists, therapy trainers, and mental health service managers. Dissemination will be via interactive 'roadshow' events with PPRG peer groups to gauge whether they feel the

Perspective	Feature	Impact	Coding
Patient	absolute words, profanity	negative	customised dictionaries
Clinician	positive sentiment	positive	valence and polarity
Linguist	first person pronouns	negative	ratio singular:plural
NLP researcher	utterance length	positive	word, character counts

Table 1: Table Examples of Candidate Language Features
Perspective: professional alignment of the core project team member suggesting the language feature.
Impact: expected relationship between the feature and level of patient engagement in the interaction.
Coding: method used to extract from the text using Python [details available from authors on request].

Observed	Predicted Negative	Predicted Neutral	Predicted Positive	Percent Correct
Negative	411	94	176	60.4%
Neutral	99	7,766	564	92.1%
Positive	1,223	3,871	11,619	69.5%
Overall Percentage	6.7	45.4	47.9	**74%**

Table 2: Initial Results for Classification of Level of Clinical Engagement

co-design process adds to the credibility of the tool.

Agile Science Approach (Hekler et al., 2016) Repeated engagement with end-users is intended to build trust (Carr, 2020) and emulates industry best practice. The project leverages specialist support from a social enterprise[2] on coproduction aspects (Hickey et al., 2018), and a digital health industry partner[3] on user experience (UX) design.

Collaborative (Machine) Learning in the tool and the process: In combination with the agile, participatory approach, the use of Human-in-the-Loop techniques will enable refinement of definitions and expose and explore tacit and latent knowledge in assessment of psychotherapy through direct involvement of domain experts in model development. Through prototyping a person-centred active learning process, we anticipate a two-way exchange of insights which will clarify what helps and what hinder the psychotherapy process.

Using evidence-based tools to capture key considerations: TrustScapes[4] were used to identify the core factors contributing to trust throughout the model pipeline (data, processing, deployment). Combined with a PROSOCIAL approach[5], this elicited fundamental stakeholder requirements for the qualities of an engaging, interactive feedback interface, and actions needed to mitigate wider concerns about its acceptability. The Software Usability Scale Plus (SUS+ (Bangor et al., 2009)) will be used as a proxy metric to evaluate explainability, supplemented by detailed qualitative feedback through 'think-aloud' exercises (Garcia et al., 2018). Measurement of trust in XAI is a new and developing field (Hoffman et al., 2019; Jacovi et al., 2020; Mohseni et al., 2020) whereas the SUS+ is well established in Human Computer Interaction, seen as a practical compromise to capturing important aspects (Davis et al., 2020) (usability and trust are interdependent (Acemyan and Kortum, 2012)), and has been used as the basis of measures of quality of explanations (Holzinger et al., 2020).

Future directions: The first PPRG workshop also started the process of gathering feedback on what is a good explanation (Danilevsky et al., 2020) and recommendation format, from each perspective (Arya et al., 2019). Over the next 3 months, we will continue to refine the classification tool, and then use it to accelerate annotation of motivation in turns-of-speech in a separate research dataset of anonymised transcripts of mainstream counselling for depression, and update the classification algorithm to increase generalisability of the tool (Topol, 2020). Given that behaviour change involves a degree of persuasion, we will explore whether we can leverage insights from Argumentation Theory to augment the model (Clos et al., 2014) and other, related developments in the field of NLP for mental

[2] Academy for Recovery Coaching CIC
[3] https://virtualhealthlabs.com/
[4] https://UnBIAS.wp.Horizon.ac.uk/fairness-toolkit/
[5] https://prosocial.world

health (e.g. unobtrusive measures of psychological inflexibility (Berkout et al., 2020), empathy (Sharma et al., 2020)). Using Natural Language Generation (NLG) for XAI (Reiter, 2019) we will test whether the model can provide its rationale in plain English matched to terms each perspective understands (Tomsett et al., 2018). We will be exploring the different use cases of justification, improvement, control and discovery (Adadi and Berrada, 2018), and investigating how the predictive ability of engagement language markers relate to those of symptomatology (Losada et al., 2019).

Acknowledgments

This work is funded by a Engineering and Physical Sciences Research Council's (EPSRC) Human Data Interaction (HDI) Network "Future of Mental Health" [6] and support from Health Data Research - UK[7].

This report is supported by the National Institute for Health Research[8] (NIHR Development and Skills Enhancement Award, Dr Sam Malins, NIHR300822). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

References

Claudia Ziegler Acemyan and Philip Kortum. 2012. The relationship between trust and usability in systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):1842–1846.

A. Adadi and M. Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques.

Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual sus scores mean: Adding an adjective rating scale. *J. Usability Studies*, 4(3):114–123.

[6] https://hdi-network.org/fmh/
[7] https://www.hdruk.ac.uk/
[8] https://www.nihr.ac.uk/

Olga V. Berkout, Angela J. Cathey, and Dmytry V. Berkout. 2020. Inflexitext: A program assessing psychological inflexibility in unstructured verbal data. *Journal of Contextual Behavioral Science*, 18:92–98.

Sarah Carr. 2020. 'ai gone mental': engagement and ethics in data-driven technology for mental health. *Journal of Mental Health*, 29(2):125–130.

Jérémie Clos, Nirmalie Wiratunga, Joemon Jose, Stewart Massie, and Guillaume Cabanac. 2014. Towards argumentative opinion mining in online discussions. In *Proceedings of the SICSA Workshop on Argument Mining (the Scottish Informatics & Computer Science Alliance)*, page 10.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing.

Brittany Davis, Maria Glenski, William Sealy, and Dustin Arendt. 2020. Measure utility, gain trust: Practical advice for xai researcher.

Jaime Delgadillo, Omar Moreea, and Wolfgang Lutz. 2016. Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy*, 79:15–22.

DHSC. 2019. Code of conduct for data-driven health and care technology.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 134. ACM.

Francisco Javier Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 99–108.

Eric B. Hekler, Predrag Klasnja, William T. Riley, Matthew P. Buman, Jennifer Huberty, Daniel E. Rivera, and Cesar A. Martin. 2016. Agile science: creating useful products for behavior change in the real world. *Translational behavioral medicine*, 6(2):317–328.

G Hickey, S Brearley, T Coldham, S Denegri, G Green, S Staniszewska, D Tembo, K Torok, and K Turner. 2018. Guidance on co-producing a research project. *Southampton: INVOLVE*.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for explainable ai: Challenges and prospects.

Andreas Holzinger, André Carrington, and Heimo Müller. 2020. Measuring the quality of explanations: The system causability scale (scs). *KI - Künstliche Intelligenz*, 34(2):193–198.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2020. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai.

Gordon V Kass. 1980. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2):119–127.

Lorenzo Lorenzo-Luaces, Robert J. DeRubeis, Annemieke van Straten, and Bea Tiemens. 2017. A prognostic index (pi) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders*, 213:78–85.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk at clef 2019: Early risk prediction on the internet (extended overview). In *CLEF (Working Notes)*.

Jay Magidson. 1994. The chaid approach to segmentation modeling: Chi-squared automatic interaction detection. In Richard P. Bagozzi, editor, *Advanced methods of marketing research*, pages 118–159. Blackwell, Cambridge, UK.

Sam Malins, Nima Moghaddam, Richard Morriss, and Thomas Schröder. 2020a. Extending the use of routine outcome monitoring: Predicting long-term outcomes in cognitive behavioral therapy for severe health anxiety. 30(5):662–674. PMID: 31438807.

Sam Malins, Nima Moghaddam, Richard Morriss, Thomas Schröder, Paula Brown, Naomi Boycott, and Chris Atha. 2020b. Patient activation in psychotherapy interactions: Developing and validating the consultation interactions coding scheme (cics). *Journal of Clinical Psychology*, 76(4):646–658.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A multidisciplinary survey and framework for design and evaluation of explainable ai systems.

Andrés Páez. 2019. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459.

Ehud Reiter. 2019. Natural language generation challenges for explainable ai.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support.

Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems.

Eric J. Topol. 2020. Welcoming new guidelines for ai clinical research. *Nature Medicine*, 26(9):1318–1320.

Shao-Yen Tseng, Brian R. Baucom, and Panayiotis G. Georgiou. 2017. Approaching human performance in behavior estimation in couples therapy using deep sentence embeddings. In *INTERSPEECH*, pages 3291–3295.

Glenn Waller and Hannah Turner. 2016. Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy*, 77:129–137.

Appendix A. Further information on CICS

A recently developed tool was deemed suitable for automation using NLP because of its focus on turn-by-turn language use in psychological therapy. The Consultation Interaction Coding Scheme (Malins et al., 2020b) offers reliable turn-by-turn assessment of interaction-types, incorporating both client and therapist responses. Using the CICS, in-session therapist-client turns-of-speech are first categorized into one of seven interaction types (action planning; evaluations of self or therapy; information discussion; noticing change or otherwise; problem description; problem analysis; structuring) and then rated -2 to +2 based on the degree of patient activation observable in the interaction. Positive ratings indicate high patient activation and engagement; negative ratings indicate low patient activation and disengagement. A series of studies have now indicated that CICS-rated psychological therapy interactions at initial sessions predict wellbeing across the course of therapy and a range of health outcomes across 12-month follow-up. This means that language features in the interactions at the first sessions of psychological therapy predicted health anxiety, generalised anxiety, depression, quality of life, general health, functioning, and somatic symptoms up to 12 months later. Specifically, if clients gave more positive evaluations of themselves or the therapy at initial sessions then better outcomes followed. Similarly, where clients were more actively engaged in structuring initial sessions and choosing session tasks, health improvements were greater. Conversely, larger proportions of initial sessions spent describing problems (as opposed to more active discussion of what might be done with problems) predicted poorer outcomes.

The Natural Language Generation Pipeline, Neural Text Generation and Explainability

Juliette Faille
CNRS/LORIA
Nancy, France
juliette.faille@loria.fr

Albert Gatt
University of Malta
Msida, Malta
albert.gatt@um.edu.mt

Claire Gardent
CNRS/LORIA
Nancy, France
claire.gardent@loria.fr

Abstract

End-to-end encoder-decoder approaches to data-to-text generation are often black boxes whose predictions are difficult to explain. Breaking up the end-to-end model into sub-modules is a natural way to address this problem. The traditional pre-neural Natural Language Generation (NLG) pipeline provides a framework for breaking up the end-to-end encoder-decoder. We survey recent papers that integrate traditional NLG sub-modules in neural approaches and analyse their explainability. Our survey is a first step towards building explainable neural NLG models.

1 Motivation

The end-to-end encoder-decoder is a popular neural approach that is efficient to generate fluent texts. However it has often been shown to face some adequacy problems such as hallucination, repetition or omission of information. As the end-to-end encoder-decoder approaches are often "black box" approaches, such adequacy problems are difficult to understand and solve.

In contrast, pre-neural NLG has often integrated a number of sub-modules implementing three main NLG sub-tasks (Reiter and Dale, 2000): macroplanning ("What to say"), microplanning and surface realisation ("How to say").

To improve adequacy and provide for more explainable approaches, recent work has proposed integrating traditional pre-neural NLG sub-modules into neural NLG models. In this paper, we survey some[1] of this work, focusing mainly on generation from data- and meaning representations[2]. Table 1 lists the approaches we consider. We start by identifying which NLG sub-tasks have been modeled in these approaches using which methods (Sec. 2-4). We then go (Sec. 5) on to briefly discuss to which extent the methods used by each of these models may facilitate explainability.

2 Macroplanning

Macroplanning is the first subtask of the traditional pre-neural NLG pipeline. It answers the "what to say" question and can be decomposed into selecting and organising the content that should be expressed in the generated text.

2.1 Content Determination

Content determination is the task of selecting information in the input data that should be expressed in the output text. The importance of this subtask depends on the goal of a generation model. In the papers surveyed, papers which verbalise RDF or Meaning Representations (MR) input do not perform content determination, while Shen et al. (2019), who generate headlines from source text, do.

In this approach, content selection is viewed as a sequence labelling task where masking binary latent variables are applied to the input. Texts are generated by first sampling from the input to decide which content to cover, then decoding by conditioning on the selected content. The proposed content selector has a ratio of selected tokens that can be adjusted, bringing controllability in the content selection.

It should also be noted that in template-based approaches such as (Wiseman et al., 2018), which use templates for text structuring (cf. Sec. 2.2), the template choice determines the structure of the output text but also has an influence on the content selection since some templates will not express some of the input information. For instance, the output 2 in

[1] Given the space limitations, the survey is clearly not exhaustive.

[2] We also include (Shen et al., 2019)'s model for text-to-text generation as it provides an interesting module for content selection which few of the papers we selected address.

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 16–21,
Dublin, Ireland, 18 December 2020.

Contribution	Content Selection	Document Structuring	REG	Input
Moryossev 19b		**Supervised**		RDF triples
Moryossev 19a		**Supervised**	**Rule-based with LM score**	RDF triples
Sha 17		**Attention**		Tables
Ferreira 19		**Supervised**	**Neural**	RDF triples
Laha 20			Rule-based	RDF triples
Gehrmann 18		**LV Template**	Coverage and length penalty	MR
Shen 20		**LV Hierarchical + Attention**		RDF triples
Shen 19	**LV**			Text
Wiseman 18	LV Template	**LV Template**		Tables
Zhao 20		**Supervised**		RDF triples
Shao 19		**LV Hierarchical**	Plan variations	Tables
Distiawan 18		**Structure encoding**		RDF triples

Table 1: Summary of the NLG models for the sub-tasks Content Selection, Document structuring and REG. The bold types indicates the main sub-task(s) modeled in each contribution and normal type the sub-task(s) that are of lesser importance in the contribution. The input type is given in the last column. LV stands for Latent Variable.

Table 2 does not include the input customer rating information.

2.2 Document structuring

Document structuring is the NLG sub-task in which the previously selected content is ordered and divided into sentences and paragraphs. The goal of this task is to produce a text plan. Many approaches choose to model document structuring. Four main types of approaches can be distinguished depending on whether the content plan is determined by latent variables, explicit content structuring, based on the input structure or guided by a dedicated attention mechanism.

Latent Variable Approaches One possible way to model content structure is to use latent variables. Wiseman et al. (2018) introduce a novel, neural parameterization of a hidden semi-markov model (HSMM) which models latent segmentations in an output sequence and jointly learns to generate. These latent segmentations can be viewed as templates where a template is a sequence of latent variables (transitions) learned by the model on the training data. Decoding (emissions) is then conditioned on both the input and the template latent variables. Intuitively, the approach learns an alignment between input tokens, latent variables and output text segments (cf. Table 2). A key feature of this approach is that this learned alignment can be used both to control (by generating from different templates) and to explain (by examining the mapping between input data and output text mediated by the latent variable) the generation model.

Similarly, Gehrmann et al. (2018) develop a mixture of models where each model learns a latent sentence template style based on a subset of the input. During generation and for each input, a weight is assigned to each model. For the same input information, two templates could produce the outputs "There is an expensive British restaurant called the Eagle" and "The Eagle is an expensive British Restaurant". The template selection defines in which order the information should be expressed and therefore acts as a plan selection.

Latent variable approaches have also been proposed for so-called hierarchical approaches where the generation of text segments, generally sentences, is conditioned on a text plan. Thus, Shen et al. (2020) propose a model where, given a set of input records, the model first selects a data record based on a transition probability which takes into account previously selected data records and second, generates tokens based on the word generation probability and attending only to the selected data record. This "strong attention" mechanism allows control of the output structure. It also reduces hallucination by using the constraints that all data records must be used only once. The model automatically learns the optimal content planning by exploring exponentially many segmentation/correspondence possibilities using the forward algorithm and is end-to-end trainable.

Similarly Shao et al. (2019) decompose text generation into a sequence of sentence generation subtasks where a planning latent variable is learned based on the encoded input data. Using this latent variable, the generation is made hierarchically with a sentence decoder and a word decoder. The plan decoder specifies the content of each output sentence. The sentence decoder also improves high-level planning of the text. Indeed this model helps capture inter-sentence dependencies in particular

Input	name[Travellers Rest Beefeater], customerRating[3 out of 5], area[riverside], near[Raja Indian Cuisine].
Output 1	[Travellers Rest Beefeater]$_{55}$ [is a]$_{59}$ [3 star]$_{43}$ [restaurant]$_{11}$ [located near]$_{25}$ [Raja Indian Cuisine]$_{40}$ [.]$_{53}$
Template 1	$z^i = \langle 55, 59, 43, 11, 25, 40, 53 \rangle$.
Output 2	[Travellers Rest Beefeater]$_{55}$ [is a]$_{59}$ place to eat]$_{12}$ [located near]$_{25}$ [Raja Indian Cuisine]$_{40}$ [.]$_{53}$
Template 2	$z^i = \langle 55, 59, 12, 25, 40, 53 \rangle$.

Table 2: Example templates and outputs segmentation from (Wiseman et al., 2018)'s approach

thanks to the global planning latent variable and attention mechanisms in the sentence decoder.

Remark. Learning a template can cover different NLG subtasks at once. For instance Gehrmann et al. (2018) use sentence templates, which determine the order in which the selected content is expressed (document structuring), define aggregation and for some cases encourage the use of referring expressions and of some turns of phrase (usually included in the lexicalisation sub-task) and defines to some extent the surface realization.

Explicit Content Structuring using Supervised Learning. Other approaches explicitly generate content plans using supervised learning.

In (Moryossef et al., 2019b), a text plan is a sequence of sentence plans where each sentence plan is an ordered tree. Linearisation is then given by a pre-order traversal of the sentence trees. The authors adopt an overgenerate-and-rank approach where the text plans are generated using symbolic methods and ranked using a product of expert model integrating different probabilities such as the relation direction probability (e.g. the probability that the triple {A, manager, B} is expressed as "A is the manager of B" or, in reverse order, as "B is managed by A") or the relation transition probability (which relations are usually expressed one after the other, e.g. birth place and birth date). Moryossef et al. (2019a) propose a variant of this model where the generation and choice of the plan to be realized is done by a neural network controller which uses random truncated DFS traversals. This new planner is achieving faster performance compared to (Moryossef et al., 2019b).

In (Castro Ferreira et al., 2019) templates are lists of ordered triples divided into sentences. Castro Ferreira et al. (2019) first order the input triples in the way they will be expressed and then divides this ordered list into sentences and paragraphs. This ordering of triples and segmentation into sentences is studied with different models : two rule-based baselines (which apply either random selection of triples or most frequent order seen on the training set) and two neural models (GRU and Transformer). They show that neural models perform better on the seen data but do not generalize well on unseen data.

Zhao et al. (2020) model a plan as a sequence of RDF properties which, before decoding, is enriched with its input subject and object. A Graph Convolutional Network (GCN) encodes the graph input and a Feed Forward Network is used to predict a plan which is then encoded by an LSTM. The LSTM decoder takes as input the hidden states from both encoders. In this approach the document structuring sub-task is tackled by an additional plan encoder.

Input structure encoding Some approaches use the structure of the input to constrain the order in which input units are verbalised. Thus, Distiawan et al. (2018) capture the inter and intra RDF triples relationships using a graph-based encoder (GRT-LSTM). It then combines topological sort and breadth-first traversal algorithms to determine in which order the vertices of the GRT-LSTM will be input with data during training thereby performing content planning.

Dedicated Attention mechanisms Instead of encoding input structure, some of the approaches use attention mechanisms to make their model focus on specific aspects of the data structure. Sha et al. (2018) take advantage of the information given by table field names and by relations between table fields. They use a dispatcher before the decoder. The dispatcher is a self-adaptative gate that combines content-based attention (on the content of the field and on the field name of the input table) and link-based attention (on the relationships between input table fields).

3 Microplanning

Microplanning is the NLG sub-task which aims at defining "how to say" the information that was selected and structured during macroplanning.

3.1 Referring Expression Generation (REG)

Few approaches explicitely model the REG subtasks. In (Moryossef et al., 2019a), REG is handled in a postprocessing step, using names for first mentions, and subsequently the pronoun or string with the highest BERT LM score. Similarly, Laha et al. (2020) use heuristic sentence compounding and coreference replacement modules as postprocessing steps. Castro Ferreira et al. (2019) explore both a the baseline model which systematically replaces delexicalised entities with their Wikipedia identifiers and the integration in the NLG pipeline of the NeuralREG model (Castro Ferreira et al., 2018). NeuralREG uses two bidirectional LSTM encoders which encode the pre- and post-contexts of the entity to be referred to. An LSTM decoder with attention mechanisms on the pre- and post-contexts generates the referring expression. Gehrmann et al. (2018) use copy-attention to fill in latent slots inside of learned templates where slots are most to be filled with named entities.

3.2 Lexicalisation

Lexicalisation maps input symbols to words. In neural approach, lexicalisation is mostly driven by the decoder which produces a distribution over the next word, from which a lexical choice is made. The copy mechanism introduced by See et al. (2017) is also widely used as it allows copying from the input (Sha et al., 2018; Moryossef et al., 2019b; Laha et al., 2020). At each decoding step, a learned "switch variable" is computed to decide whether the next word should be generated by the S2S model or simply copied from the input. Inspecting the value of the switch variable permits assessing how much lexicalisation tends to copy vs to generate and can provide some explainability in the lexicalisation sub-task. Finally, a few approaches use lexicons and rule-based mapping. In particular, Castro Ferreira et al. (2019) use a rule-based model to generate the verbalization of RDF properties.

4 Surface realisation

Surface realisation is the last NLG task and consists in creating a syntactically well-formed text out of the representations produced by the previous step. While surface realisation is at the heart of generation when generating from meaning representations, it is largely uncharted in data- and table-to-text NLG and results either from the decoder language model (which decides on the words and thereby indirectly on the syntax of the generated text) or from the templates used for generation (Castro Ferreira et al., 2019; Moryossef et al., 2019b; Wiseman et al., 2018).

5 Conclusion

Explainable models enable a clear understanding of how the output generated by the model relates to its input. In this short paper, we surveyed a number of neural data-to-text generation models which implement some or all of the NLG pipeline sub-tasks with the aim of identifying methods which could help enhance explainability in neural NLG.

Our survey highlights two main ways of enhancing explainability: explicit intermediate structures produced by neural modules modeling the NLG pipeline subtasks or latent variables modeling the interface between these modules.

Thus (Castro Ferreira et al., 2019)'s supervised pipeline model outputs content plans, sentence templates and referring expressions which can all be examined, quantified and analysed thereby supporting a detailed qualitative analysis of each subtasks. Similarly, Moryossef et al. (2019b,a) output explicit text plans and text plan linearisations and Zhao et al. (2020) text plans.

In contrast, the models introduced in (Shao et al., 2019; Wiseman et al., 2018; Gehrmann et al., 2018; Shen et al., 2019, 2020) are based on latent variables which mediate the relation between input and output tokens and intuitively, model a document plan by mapping e.g., input RDF triples to text fragments. As illustrated in Table 2 which shows examples of latent templates used to generate from the input, latent variables provide a natural means to explain the model's behaviour i.e., to understand which part of the input licenses which part of the output. They are also domain agnostic and, in contrast to the explicit pipeline models mentioned in the previous paragraph, they do not require the additional creation of labelled data which often relies on complex, domain specific, heuristics.

A third alternative way to support explainability is model analysis such as supported e.g., by the AllenNLP Interpret toolkit (Wallace et al., 2019) which provides two alternative means for interpreting neural models. Gradient-based methods explain a model's prediction by identifying the importance of input tokens based on the gradient of the loss with respect to the tokens (Simonyan et al., 2014)

while adversarial attacks highlight a model's capabilities by selectively modifying the input.

In future work, we plan to investigate whether domain agnostic, linguistically inspired intermediate structures such as meaning representations could be used to both support explainability and improve performance. Another interesting direction for further research would be to develop common evaluation benchmarks and metrics to enable a detailed analysis and interpretation of how neural NLG models perform for each of the NLG pipeline sub-tasks. Finally, while most of the approaches we surveyed concentrate on modeling the interaction between content planning and micro-planning, it would be useful to investigate whether any of the methods highlighted in this paper could be exploited to explore and improve the explainability of the various micro-planning sub-tasks (lexicalisation, aggregation, regular expression generation, surface realisation).

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. Research reported in this publication is part of the project NL4XAI. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains.

References

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Bayu Distiawan, Jianzhong Qi, Rui Zhang, and Wei Wang. 2018. Gtr-lstm: A triple encoder for sentence generation from rdf data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1627–1637.

Sebastian Gehrmann, Falcon Dai, Henry Elder, and Alexander Rush. 2018. End-to-end content and plan selection for data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 46–56, Tilburg University, The Netherlands. Association for Computational Linguistics.

Anirban Laha, Parag Jain, Abhijit Mishra, and Karthik Sankaranarayanan. 2020. Scalable micro-planned generation of discourse from structured data. *Computational Linguistics*, 45(4):737–763.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019a. Improving quality and efficiency in plan-based neural data-to-text generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 377–382, Tokyo, Japan. Association for Computational Linguistics.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019b. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. 2018. Order-planning neural text generation from structured data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5414–5421. AAAI Press.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text

generation with planning-based hierarchical variational model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3257–3268, Hong Kong, China. Association for Computational Linguistics.

Xiaoyu Shen, Ernie Chang, Hui Su, Cheng Niu, and Dietrich Klakow. 2020. Neural data-to-text generation via jointly learning the segmentation and correspondence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7155–7165, Online. Association for Computational Linguistics.

Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. 2019. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590, Hong Kong, China. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, Banff, Canada.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. In *EMNLP*, pages 7–12, Hong Kong, China.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, volume 1.

Towards Harnessing Natural Language Generation to Explain Black-box Models

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

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

Abstract

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

1 Introduction

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

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

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

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

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

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

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

2 Open Challenges in XAI

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

2.1 Designing Explainable Models

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

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

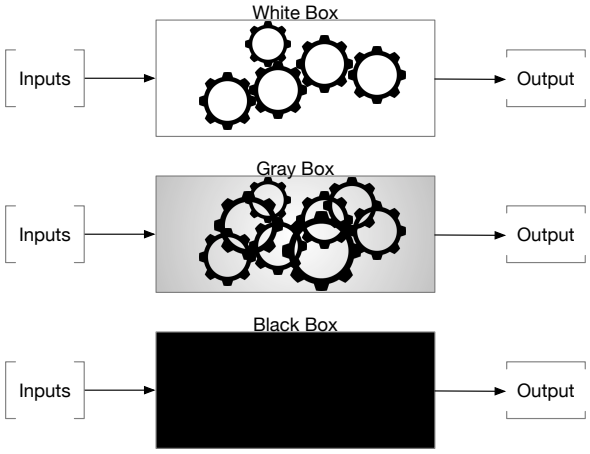

Figure 1: A block diagram representation of different models. White boxes have a clear and decomposable internal representation and processing. Gray boxes are still decomposable but understanding them is less straightforward. For black boxes, processing is assumed unknown and only the input-output behaviour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

2.1.1 White-box Models

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

2.1.2 Gray-box Models

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

2.1.3 Black-box Models

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

2.2 Implementing Explanation Interfaces

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

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

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

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

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

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

2.3 Evaluating Explanation Systems

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

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

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

3 Explaining with Natural Language

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

3.1 Template-based Generation

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

times produces sentences that are non-natural and lack variation.

3.2 End-to-end Generation

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

3.3 Evaluating Natural Language Generation

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

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

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

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

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

4 Conclusions

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

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

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

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

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Abstract

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

1 Introduction

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

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

2 Bayesian Networks

2.1 Overview

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

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

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

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

2.2 What can be Explained?

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

- Explanation of the evidence typically amounts to providing the most probable explanation of a node of interest in the network by selecting the configurations of variables that are most likely to have resulted in the available evidence. In BNs this is often done by calculating the maximum a-posteriori probability for the evidence. This can aid in situations such as medical diagnoses and legal cases.

- Explanation of the model involves describing the structure of the network and the relationships contained within it. Unlike other discriminative models such as decision trees, prior probabilities and expert knowledge may have been used to construct the BN and may need to be explained. This can be used to provide domain knowledge for end users or for debugging a model.

- Explanation of the reasoning has the goal of describing the reasoning process in the network which took place to obtain a result. This can also include explanations of why a certain result was not obtained, or counterfactual explanations about results that could be obtained in hypothetical situations (Constantinou et al., 2016).

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

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

3 Need for Natural Language Explanation

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

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

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

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

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

4 Textual Explanations of BNs

4.1 State of the Art

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

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

Keppens (2019) also described an approach to the determination of content from a BN as part of the NLG pipeline, using the support graph method described by Timmer et al. (2017). It is then shown how this content is trimmed and ordered at the high-

level planning stage. In order to implement the high level-plan, sentence structures are generated at the micro-planning stage.

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

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

4.2 Discussion and Challenges for Future Work

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

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

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

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

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

- Past work on the linguistic expression of probabilistic values is often not considered. Devel-

opers commonly opt to print numerical values leading to less acceptable explanations.

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

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

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

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

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

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

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

5 Conclusion

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

Acknowledgments

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Explaining data using causal Bayesian networks

Jaime Sevilla
Aberdeen University
j.sevilla.20@abdn.ac.uk

Abstract

We introduce Causal Bayesian Networks as a formalism for representing and explaining probabilistic causal relations, review the state of the art on learning Causal Bayesian Networks and suggest and illustrate a research avenue for studying pairwise identification of causal relations inspired by graphical causality criteria.

1 From Bayesian networks to Causal Graphical Models

Bayesian networks (BNs) are a class of probabilistic graphical models, originally conceived as efficient representations of joint probability distributions.

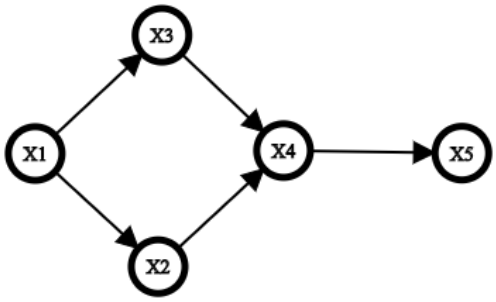

Figure 1: Bayesian network representing the probability distribution $P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1), P(x_4|x_2, x_3)P(x_5|x_4)$

A great deal of work has been dedicated in the last decades to understanding how to represent knowledge as BNs, how to perform efficient inference with BNs and how to learn BNs from data (Koller and Friedman, 2009).

Despite having been overshadowed by subsymbolic approaches, BNs are attractive because of their flexibility, modularity and straightforward statistical interpretation.

On top of that, BNs have a natural interpretation in terms of causal relations. Human-constructed BNs tend to have arrows whose directionality respects the causal intuitions of their architects.

Furthermore, recent work has extended Bayesian Networks with causal meaning (Pearl, 2009; Spirtes et al., 2001). The result are Causal Bayesian Networks and Causal Structural Models, that ascribe new meaning to BNs and extend classical inference with new causal inference tasks such as interventions (eg will the floor get wet if we turn the sprinkler on?) and counterfactuals (eg would this person have received a good credit rating if they had a stable job?).

In this paper we will review work on the area of using Bayesian networks to model causal relationship, and consider one future research direction to explore, concerning the identification of the causal link between pairs of variables.

2 Learning Causal Bayesian Networks

Considerations of causality also affect how Bayesian Networks should be learnt. Manually built Bayesian networks usually respect our causal intuitions. But Bayesian networks learnt from data may not respect the underlying causal structure that generated the data.

Indeed, each probability distribution can be represented by several different Bayesian Networks - and we can group Bayesian Networks graphs in classes capable of representing the same probability distributions, their Markov equivalence class.

Traditional BN learning methods such as score maximization (Cussens et al., 2017) cannot distinguish between members of the same Markov equivalence class, and will be biased towards outputting a Bayesian structure that fits the data well but does not necessarily match the underlying causal mechanisms.

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NLXAI 2020), pages 34–38,
Dublin, Ireland, 18 December 2020.

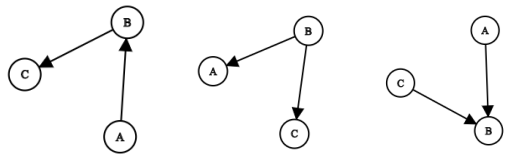

Figure 2: Three Bayesian Networks. The left and the middle one are Markov-equivalent, but the third one isn't equivalent to the other two - in fact, the left one is the only member of its reference class. Hence, if the data is compatible with the right BN and there are no latent variables BN we will be able to conclude that A causes B. But if the data is compatible with the left (and therefore the middle) BN then the orientation of the edge $A - B$ is arbitrary, and we cannot infer just from the data the causal relations between the variables.

This is a key problem for explaining the outcome of Bayesian Network learning algorithms. Experts usually avoid altogether a causal language - instead framing their explanations in terms of association. But we would like to be able to actually explain when a relation is causal and when we do not have enough information to tell one way or another.

In order to do this, we need our learning methods to distinguish and explain when their edge orientation decisions are arbitrary (ie there is another BN compatible with the data [1]where the edge is oriented in a different way) or necessary (ie the edge has this orientation in every diagram compatible with the data we have) - since only in the latter situation can we guarantee that the orientation will respect causality.

3 Previous work

This problem of causal discovery based on graphical models is reviewed in depth in (Glymour et al., 2019). In this article the authors introduce three families of causal discovery algorithms:

- Constraint based algorithms that rely on conditional independence tests to orient the edges in a graph. See for example the PC algorithm (Spirtes et al., 2001).
- Score based algorithms that greedily optimize a score function to orient the edges in a graph. See for example the Greedy Equivalence Search (Chickering, 2002).
- Functional algorithms that use stronger assumptions about the relation between two directly related variables to distinguish cause and effect. See for example the post-nonlinear causal model (Zhang and Hyvarinen, 2009).

The problem is considerably more difficult when we allow the possibility of unmeasured ('latent') common causes of the variables in our dataset.

This situation is arguably more representative of usual datasets, and requires specialized methods to be addressed. (Zhang, 2008) proposed a constraint-based learning algorithm that is provably sound and complete, assuming correct conditional independence decisions. The algorithm was later refined in (Claassen and Heskes, 2011).

4 A graphical test of causality and missing confounders

However, J. Zhang's and similar methods rely on frequentist and high order conditional independence tests to learn the causal structure, which are prone to error. The serial nature of the algorithm means that early errors in the conditional independence decisions lead to more errors later.

Ideally, we would like to have our methods of learning causality from observational data be more robust to statistical noise, and do not let errors propagate through the graph.

This is especially important when we are not interested in learning the complete structure of the graph, but rather we want to study the particular relation between a variable we could manipulate (the 'exposure') and a variable we care about (the 'outcome').

This problem has been discussed in depth in the context of econometrics, where structural equation modelling (Kaplan, 2020) and instrumental variable estimation methods (Reiersöl, 1945) are widely used tools for causal inference.

While structural equation modelling provides satisfactory answers to many questions of causal estimation, they are hard to interpret and use. Graphical models could lead us to better explanations of

[1]We have glossed over what compatible exactly means. A necessary condition is that all the independence conditions represented via d-separation in the graph are present in the joint probability distribution of the data (Spirtes, 1996). We usually also require the reverse, that all conditional independencies in the joint probability distribution are represented via d-separation in the graph - this is called the faithfulness assumption. The faithfulness assumption renders conditional independence and d-separation effectively equivalent, and restricts the output of the algorithm to a single Markov equivalence class. A justification of why we should expect our data to be faithful to the underlying model can be found in (Pearl, 2009, Chapter 2).

the models of causality used in econometrics and other contexts. For example, instead of providing models as mathematical equations, the causality we can infer from the data could be represented graphically, and described via text using similar techniques to those that apply to explaining Bayesian Networks with natural language generation techniques (see (Reiter, 2019) for discussion).

In particular, under certain conditions we can use insights derived from causal discovery in graphical models to test conditions usually taken on faith.

For example, if we identify two additional variables Z, W and a context $S = s$ such that:

- A and B are conditionally dependent given $S = s$
- Z and W are conditionally independent given $S = s$, but are conditionally dependent given $S = s$ and $A = a$ for some value a
- Z and B are conditionally dependent given $S = s$, but conditionally independent given $S = s$ and $A = a$ for every value a

then lemma 1 from (Claassen and Heskes, 2011) implies under mild assumptions that there is a directed path from A to B in every causal bayesian network compatible with the data we have observed.

To ground this example, let's suppose that we are interested in studying the effect of a drug (A) on the health of a patient (B). We furthermore have access to information about the patient's income (Z) and whether they have health insurance (W). We also have access to a set of background information variables (O) like for example age and gender.

We assume that the causal relationships between the variables can be represented as an acyclic graphical model.

We check that the income (Z) and the drug (A) are independent conditional on some of the background variables ($S \subset O$), but dependent when we condition on $S \cup \{A\}$.

Then we check that the income (Z) and the patient's health outcome (B) are conditionally dependent given the same subset of background variables S, but independent when we condition on the drug A.

Then we will be able to assert that no matter what the true acyclic causal diagram is, there will always be a causal path that starts in the treatment (A) and ends in the patient's health outcome (B). This guarantee holds as long as we can guarantee acyclity - even if there are unmeasured latent variables in the true causal diagram.

Hence it would be appropriate to describe the data as providing evidence for the natural language explanation "the drug has an effect on the health of the patient". Note that we can only provide this explanations because of our explicit causal analysis. A traditional Bayesian analysis would only be able to conclude that the drug and the health outcome are somehow related - but it would not have been able to distinguish the direction of causation (perhaps sicker patients are more likely to be treated with the new drug!) or rule out confounding common causes (perhaps richer patients are both more likely to receive the treatment and have better health outcomes for reasons unrelated to the drug!).

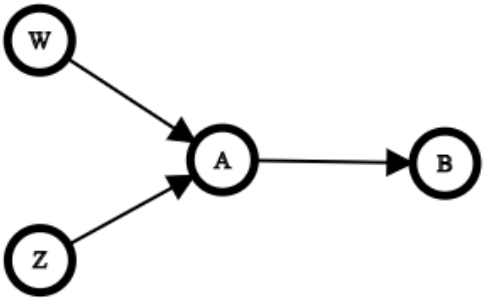

Figure 3: If the underlying causal structure follows this diagram, then because of d-separation properties we will be able to conclude that A, B, Z, W and $S = \emptyset$ satisfy the conditions we listed. Hence, every Bayesian network in the Markov-equivalence class of the diagram (including diagrams with latent variables) includes a directed path from A to B. So we will be able to unequivocally conclude that A causes B.

Like J. Zhang's causal discovery algorithm, this criteria allows the possibility of latent common causes. Unlike J. Zhang's, this criteria only depends on locally testing the conditional independence relations between A, B, Z, W, S to conclude that A is a cause of B. A similar approach is considered in (Mani et al., 2012), though in the context of global structure learning.

From an econometric perspective, the interest of the criteria above is that this condition provides a graphical test for causality and missing confounders, under the assumption of no cyclical causal relations. In particular, the conditions outlined above imply that $S = s$ blocks all confounding paths that spuriously relate A and B but blocks

no causal path between the variables. Thus if we found that the criteria holds we would be able to use standard tools such as ordinary least squares regression to quantify the strength of the causal relation $A \to B$.

Related tests already exist in the literature - for example the overidentification J-test for testing the exogeneity of instrumental variables (Stock and Watson, 2011, Section 12.3), and selection of observables for testing the sensibility of an estimate to hidden confounders (Altonji et al., 2000).

Understanding the relationship between these traditional tests and the tests derived from the graphical criteria will be an interesting multidisciplinary exercise.

5 Conclusion and next steps

In summary, the development of a local causal criteria will give us a powerful tool to build causal explanations of data, that under certain conditions can distinguish the direction of causality and quantify the strength of the underlying causal relation.

The development of this criteria will be of great help to fields eager to extract causal conclusions from historical data. For example, this could help medics and patients gain an understanding of how much of a difference a treatment would make based on the history of past patients, so they can make an informed decision about it.

It is unclear how to generalize these conditions to cover more cases and possible causal relations, how often these conditions are met, how reliable procedures of proving causality based on this type of criteria would be and how to deal with possibly contradictory evidence of causality.

Our intention is to explore these questions through our work. This will involve three avenues of research:

- Formulating and formally studying criteria for proving causal relations through mathematical definitions and proofs
- Developing my own algorithms of causal discovery based on such criteria and refining them by evaluating them on synthetic data
- Testing the performance the resulting algorithms on real datasets

We do not expect this work to be easy.

Specially challenging in the context of econometrics will be the validation of the methods used. Only seldom do we have direct experimental evidence of the causal relations in a economic domain. Because of this, initial experimentation should focus on explaining observational data in domains where there is a strong and well-established theory of causation, such as price-demand modelling.

Another key difficulty is the requirement of conditional independencies - it will often be impossible in econometric contexts to render variables conditionally independent. Thus part of our work will require us to relax the conditions of Y-structure based causal discovery to exploit weaker forms of conditional independence. For example, we could look into interaction information (McGill, 1954) or related concepts from information theory.

Finally, there is a problem on explaining this graphical reasoning to users. It is not obvious why Y-structures imply a causal relationship. It may be fruitful to draw an analogue between this method and how humans infer causation, to make them more intuitive.

We believe that this work will help us better understand how to study causal relationships from observational data, which will have long reaching applications in econometrics, medicine and other fields of practice that routinely need to rely on observational data for their analyses.

Furthermore, causal graphical models have an advantage compared to black box learning and reasoning models due to their ability to address causal queries. This could be leveraged to marginally push the field of AI towards methods inspired by probabilistic graphical models, which are arguably more transparent and will facilitate goal alignment.

Acknowledgements

I thank my supervisors Ehud Reiter and Nava Tintarev for thorough discussion and support.

I also thank the anonymous reviewers for the NL4XAI for kindly providing constructive feedback to improve the paper.

This research has been supported by the NL4XAI project, which is funded under the European Union's Horizon 2020 programme, grant agreement 860621.

References

Joseph G Altonji, Todd E Elder, and Christopher R Taber. 2000. Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic

Schools. Working Paper 7831, National Bureau of Economic Research. Series: Working Paper Series.

David Maxwell Chickering. 2002. Optimal Structure Identification With Greedy Search. *Journal of Machine Learning Research*, 3(Nov):507–554.

Tom Claassen and Tom Heskes. 2011. A structure independent algorithm for causal discovery. In *In ESANN'11*, pages 309–314.

James Cussens, Matti Järvisalo, Janne H. Korhonen, and Mark Bartlett. 2017. Bayesian Network Structure Learning with Integer Programming: Polytopes, Facets and Complexity. *Journal of Artificial Intelligence Research*, 58:185–229.

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10. Publisher: Frontiers.

David Kaplan. 2020. *Structural Equation Modeling (2nd ed.): Foundations and Extensions*, 2nd edition. Thousand Oaks, California.

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Subramani Mani, Peter L. Spirtes, and Gregory F. Cooper. 2012. A theoretical study of Y structures for causal discovery. *arXiv:1206.6853 [cs, stat]*. ArXiv: 1206.6853.

William J. McGill. 1954. Multivariate information transmission. *Psychometrika*, 19(2):97–116.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition edition. Cambridge University Press, Cambridge, U.K. ; New York.

Olav Reiersöl. 1945. Confluence analysis by means of instrumental sets of variables.

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

Peter Spirtes. 1996. Using d-separation to calculate zero partial correlations in linear models with correlated errors. Publisher: Carnegie Mellon University.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search, 2nd Edition*, volume 1. The MIT Press. Publication Title: MIT Press Books.

James Stock and Mark Watson. 2011. *Introduction to Econometrics (3rd edition)*. Addison Wesley Longman.

Jiji Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896.

Kun Zhang and Aapo Hyvarinen. 2009. On the Identifiability of the Post-Nonlinear Causal Model. page 9.

Towards Generating Effective Explanations of Logical Formulas: Challenges and Strategies

Alexandra Mayn and **Kees van Deemter**
Utrecht University
Department of Information and Computing Sciences
{a.mayn, c.j.vandeemter}@uu.nl

Abstract

While the problem of natural language generation from logical formulas has a long tradition, little attention has been paid to ensuring that the generated explanations are optimally helpful to the user. We discuss issues related to deciding what such output should look like and strategies for addressing those issues. We stress the importance of informing generation of NL explanations of logical formulas with reader studies and findings on the comprehension of logic from pragmatics and cognitive science. We illustrate the issues and potential ways of addressing them using a simple demo system's output generated from a propositional logic formula.

1 Introduction

The task of generating natural language text from logical form has a long and diverse tradition (Wang (1980), Appelt (1987), Shieber et al. (1990), to name a few early examples). It has been approached from a variety of perspectives targeting different use cases, including providing feedback to students of logic (Flickinger (2016)), users of logistic software (Kutlak and van Deemter (2015)), and explaining the output of a reasoning engine (Coppock and Baxter (2009)).

However, so far in this domain very little attention has been paid to generating output which is optimally helpful to the user, presumably a non-expert with little knowledge of formal logic, We aim to build a system which, given a logical formula, will produce an effective natural language explanation. To that end, we discuss challenges which might arise when building and scaling up such a generation system and strategies for addressing these challenges with the user in mind. We aim to conduct studies with potential users (students of logic and/or users of software which operates using logic) to determine what kinds of explanations the system should generate.

The rest of this paper is organized as follows: Section 2 summarizes related work. Section 3 discusses challenges and possibilities related to determining and ensuring effectiveness of the generated output for the users. Section 4 illustrates our points by means of a case study on producing explanations of a propositional logic formula. Section 5 concludes.

2 Related work

There have been a number of works aimed at generating text from logical form, using either rule-based (Shieber et al. (1990); Ranta (2011); De Roeck and Lowden (1986)) or statistical (Basile (2015); Zettlemoyer and Collins (2012); Lu and Ng (2011)) methods. However, only a few of them explicitly discuss the issues related to the comprehensibility and effectiveness of the generated output. De Roeck and Lowden (1986) opt for using indentation as opposed to linear text to minimize ambiguity of the generated text, while Ranta (2011)'s solution involves bulleted lists. Flickinger (2016) addresses a related issue, that of generating multiple paraphrases for a logical formula, with a view to subsequently selecting the best one - as many as almost 4500 paraphrases are generated for one formula, but the issue of filtering out ambiguous paraphrases and selecting the best one is left to future work. Kutlak and van Deemter (2015) apply transformations at the logic level with the aim of making the formula more suitable for generation.

Studies with human participants to determine what output of NLG systems is preferable have been conducted in other domains. Eugenio et al. (2005) study the effect of aggregation in the output of an automated tutoring system on the learner and find that what they call functional aggregation,

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NLXAI 2020), pages 39–43,
Dublin, Ireland, 18 December 2020.

which produces summaries of text relating to the same object or function, benefits the user more than purely syntactic aggregation does. Khan et al. (2012) conduct reading studies to investigate comprehension of referring expressions with a view to improving an NLG system's output. They investigate the interaction between brevity and clarity, and find that brevity is preferred when all alternatives are unambiguous, and otherwise a longer but unambiguous alternative is preferred, which goes to show that there is an advantage to non-brief referring expressions in terms of comprehension. Khan et al. (2012)'s NLG algorithms thus incorporate a trade-off between clarity and brevity.

3 User-oriented explanations: Challenges and strategies

There are a number of questions which need to be answered when developing an explanation-producing system aimed at making the explanations maximally helpful to the user.

As Kutlak and van Deemter (2015) point out, it is not always the case that the inputted logical formula is in a form suitable for generation. Some formulas are unnecessarily complex and would therefore tend to produce unnecessarily complex text unless optimised first. To make matters trickier, for expressive logics it is often not decidable whether two logical formulas are equivalent to each other (termed "the problem of *logical form equivalence*" by Shieber (1993)), so heuristics need to be developed to decide what transformations to apply, and how many, and to determine how suitable a formula is for generation. Kutlak and van Deemter (2015) assume as a rule of thumb that transformations which will make a formula shorter are likely to also make it easier to comprehend. However, there are some cases where making the formula longer might be warranted, e.g. if that results in a clearer NL explanation. We believe that it would be beneficial to conduct empirical studies on comprehension and preference between text variants generated based on several equivalent formulas in order to develop such heuristics.

At the NL generation stage, there are important decisions to be made as well. Which phrasings should be used and which ones should be avoided? One of the aspects which can make generation from logic challenging is that the meaning of logical connectives is not always the same as that of their natural language counterparts (Grice (1975), Moeschler (2017)). For instance, Geis and Zwicky (1971) argue that an NL conditional is often used as a logical biconditional (for example, *If you go to the party, I'll go too* is understood to imply that *if you do not go, neither will I*), while Barker-Plummer et al. (2008) show that students of logic particularly struggle with the expression of the biconditional as *just in case* because it has a very different meaning in everyday natural language.

In terms of the form of the generated text, there are a number of alternatives which have been used - linear text, bulleted lists and indentation; these presentation decisions will have an effect on the comprehensibility of the generated output. Related, what is the optimal amount of aggregation in this context? Are there situations where it is preferable not to aggregate? We argue that these questions should be addressed through controlled user experiments where reading comprehension and speed for alternatives is compared along each of these dimensions.

We also believe that such resources as the Grade Grinder Corpus (Barker-Plummer et al., 2011), which contains students' attempts to convert natural language text to FOL, can also inform us about which natural language wordings are effective and which ones should be avoided by the generator. Both number and nature of incorrect attempts by the students can be used in gaining insights as to what realizations of connectives tend to be misunderstood and what they are misunderstood as. For instance (Barker-Plummer et al., 2008) find that many errors are made when formalizing a sentence in FOL requires reordering the antecedent and the consequent.

As has been pointed out in related work (Ranta (2011); Flickinger (2016)), *ambiguity* is a challenging aspect of this generation problem: if not controlled for, bracketing or negation scope ambiguity is likely to emerge. Ranta (2011) proposes using a parser test to determine whether the generated output can have multiple readings, and select an unambiguous one that way. We believe that that is an effective solution to the ambiguity problem. However, we can imagine a case where, for a sufficiently complex formula, the generator might only produce explanations with multiple readings, or the unambiguous variant is too clunky and difficult to read. In that case, it would be beneficial to know about the respective likelihood of the alternative readings for the user. It could be, for instance, that

a reading is identified by the parser which a human is unlikely to consider. With this likelihood information, one could, for instance, select an output which has the fewest possible readings, or only one likely reading. We intend to explore whether such an approximation of likelihood can be obtained using probabilistic parsing (Jurafsky and Martin, 2009, Chapter 14).

4 Case Study: Generation from Propositional Logic

We illustrate the above points by means of a concrete example. As a starting point, we built a simple system, which takes a propositional logic formula as input, parses it into a tree structure, optionally applies transformations to the tree, and realizes the output by reading off the tree, left to right. We chose propositional logic as a base case because it is one of the simplest logics in which many of the important discussed in Section 3 emerge.

Consider the following formula, which involves the block language from Tarski's World, a software component for teaching logic to students (Barwise et al., 2000):

(1) $\neg Cube(x) \wedge \neg(Smaller(x, y) \vee SameShape(x, y))$

At the tree level, we apply a number of meaning-preserving transformations. For any formula, there is an infinite set of formulas equivalent to it, therefore heuristics need to be developed as to what makes a formula a promising candidate for generation. For simplicity, we start with a set of 8 formulas equivalent to (1), obtained by distributing negation, applying De Morgan's laws, or reversing the order of the conjuncts and disjuncts. We pass each of these formulas to a simple generator, obtaining 8 wordings. For example, (1) is realized as:

(a) x is not a cube and it is not the case that x is smaller than y or x is the same shape as y.

Which, if any, of the generated versions should be the final output? At this stage, we run a syntactic parser on the generated text to identify how many and what kind of possible readings the generated text may have. We then determine how many distinct parses each of the texts has by computing which of the parses are equivalent to each other. We find that some generated text variants come out as unambiguous, while others have as many as 11 distinct parses.

We are not quite done yet. It is worth pointing out that the ambiguities which the parser detects might not perfectly predict what ambiguities might arise for people. For example, (a) can be parsed three ways, with *it is not the case* having either narrow or wide scope. However, one could argue that the narrow scope reading is a lot less likely. That could be determined using probabilistic parsing.

Conversely, there could emerge certain mirage ambiguities, where a sentence which is grammatically unambiguous could still be understood multiple ways by the reader, e.g. we can imagine *it is not the case that x is smaller than y or the same shape as y* being misunderstood as $\neg Smaller(x, y) \vee SameShape(x, y)$ (narrow scope of negation). Such cases seem more difficult to foresee. Reader studies could be helpful in gaining insight; complexity heuristics could also be introduced with the hope that less complex sentences would be less likely to give rise to such problems.

Besides ambiguity, there is another dimension along which the effectiveness of the generated text will vary: readability. Text length and naturalness both affect readability. Interestingly, ambiguity also interacts with readability: there is evidence that ambiguous sentences are processed faster than disambiguated ones, but only when the readers do not anticipate the need to answer in-detail questions about the read text (Swets et al., 2008). Methods like aggregation may be employed to improve readability. So, (1) could also be worded as:

(b) x is not a cube, nor is it smaller than or the same shape than y.

That phrasing is also no longer ambiguous, which illustrates that aggregation can have an impact on clarity as well as readability.

Of course, a yet more natural phrasing of (1) is as follows:

(c) x is not a cube. it is at least as large as y and has a different shape.

It would be challenging to generate (c) automatically given an input formula like (1) since the system would need to have information about what *not smaller* means. Kutlak and van Deemter (2015) allow the user to enter background axioms, which partially addresses the problem: the user would

need to explicitly indicate the equivalence between *not smaller* and *at least as large*, and between *not the same* and *different shape*, in order for such output to be generated.

In the above, we have experimented with an approach that computes many possible interpretations of each candidate NL text. An interesting avenue for further research is to investigate how such a brute-force approach may be approximated by a set of heuristics, which could then be used in an approach similar to the *revision process* for NLG (Inui et al., 1992) to avoid unnecessary computation: generating output which is estimated to be the best (i.e., most clear and natural), checking these constraints and repeating the process for the next best output if one of the constraints is violated. An open and challenging question is that of generality: if we identify a set of heuristics for a certain class of formulas, how well will they generalize to a different class of formulas or set of predicates? We aim to explore that through controlled experiments.

5 Conclusion

In this paper, we addressed an aspect of the design of an NLG system for explaining the meaning of logical formulas which has often been overlooked: the needs of the user. We discussed questions which we aim to answer when building such a system, such as logical simplifications, paraphrasing and ambiguity, and considered ways in which they can be informed: reading studies with potential users, work with corpora, and insights from cognitive science and pragmatics. We illustrated these questions and potential solutions by means of an example of generating text from a propositional logic formula.

6 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 860621.

References

Douglas Appelt. 1987. Bidirectional grammars and the design of natural language generation systems. In *Theoretical Issues in Natural Language Processing 3*.

Dave Barker-Plummer, Richard Cox, and Robert Dale. 2011. Student translations of natural language into logic: The grade grinder corpus release 1.0. In *Proceedings of the 4th international conference on educational data mining*, pages 51–60.

Dave Barker-Plummer, Richard Cox, Robert Dale, and John Etchemendy. 2008. An empirical study of errors in translating natural language into logic. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Jon Barwise, John Etchemendy, Gerard Allwein, Dave Barker-Plummer, and Albert Liu. 2000. *Language, proof and logic*. CSLI publications.

Valerio Basile. 2015. From logic to language: Natural language generation from logical forms.

Elizabeth Coppock and David Baxter. 2009. A translation from logic to english with dynamic semantics. In *JSAI International Symposium on Artificial Intelligence*, pages 197–216. Springer.

Anne De Roeck and Barry GT Lowden. 1986. Generating english paraphrases from formal relational calculus expressions. In *Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics*.

Barbara Di Eugenio, Davide Fossati, Dan Yu, Susan Haller, and Michael Glass. 2005. Aggregation improves learning: experiments in natural language generation for intelligent tutoring systems. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics.

Dan Flickinger. 2016. Generating english paraphrases from logic. *From Semantics to Dialectometry*, pages 99–107.

Michael L Geis and Arnold M Zwicky. 1971. On invited inferences. *Linguistic inquiry*, 2(4):561–566.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1992. Text revision: A model and its implementation. In *International Workshop on Natural Language Generation*, pages 215–230. Springer.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Imtiaz H Khan, Kees van Deemter, and Graeme Ritchie. 2012. Managing ambiguity in reference generation: the role of surface structure. *Topics in Cognitive science*, 4(2):211–231.

Roman Kutlak and Kees van Deemter. 2015. Generating succinct english text from fol formulae. In *Procs. of First Scottish Workshop on Data-to-Text Generation*.

Wei Lu and Hwee Tou Ng. 2011. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1611–1622.

Jacques Moeschler. 2017. What logic tells us about natural language1. *The Routledge handbook of pragmatics*, page 241.

Aarne Ranta. 2011. Translating between language and logic: what is easy and what is difficult. In *International Conference on Automated Deduction*, pages 5–25. Springer.

Stuart Shieber. 1993. The problem of logical-form equivalence. *Computational Linguistics*.

Stuart Shieber, Gertjan Van Noord, Fernando CN Pereira, and Robert C Moore. 1990. Semantic-head-driven generation. *Computational Linguistics*.

Benjamin Swets, Timothy Desmet, Charles Clifton, and Fernanda Ferreira. 2008. Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36(1):201–216.

Juen-tin Wang. 1980. On computational sentence generation from logical form. In *COLING 1980 Volume 1: The 8th International Conference on Computational Linguistics*.

Luke S Zettlemoyer and Michael Collins. 2012. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*.

Argumentation Theoretical Frameworks for Explainable Artificial Intelligence

Martijn H. Demollin*
Laboratory of The New Ethos,
Faculty of Administration and Social Sciences,
Warsaw University of Technology, Poland
Martijn.Demollin@pw.edu.pl

Qurat-ul-ain Shaheen*
IIIA-CSIC, Spain
qurat@iiia.csic.es

Katarzyna Budzynska
Laboratory of The New Ethos,
Faculty of Administration and Social Sciences,
Warsaw University of Technology, Poland
Katarzyna.Budzynska@pw.edu.pl

Carles Sierra
IIIA-CSIC, Spain
sierra@iiia.csic.es

Abstract

This paper discusses four major argumentation theoretical frameworks with respect to their use in support of explainable artificial intelligence (XAI). We consider these frameworks as useful tools for both system-centred and user-centred XAI. The former is concerned with the generation of explanations for decisions taken by AI systems, while the latter is concerned with the way explanations are given to users and received by them.

1 Introduction

The enforcement of GDPR (https://gdpr-info.eu/) by the EU has made eXplainable Artificial Intelligence (XAI) into a rapidly growing area of research over the last two years. While there is no standard definition of explainable AI systems yet, the need itself is undisputed as evidenced by the GDPR requirements. Also, there is agreement that explainability for AI systems is as diverse as the systems themselves. Neerincx et al. have defined three phases in the explanation of an AI system: (1) explanation generation, (2) explanation communication, and (3) explanation reception (Neerincx et al., 2018). Based on this, recent XAI literature can be divided into two types: *system-centred* and *user-centred* XAI.

System-centred XAI is focused on phase 1. Broadly, systems fall into two main categories: black-box subsymbolic systems such as those based on deep learning and white-box symbolic systems like decision trees or rule-based. A consequence of the GDPR implementation has been a recent explosion in grey-box systems, which aim to add some symbolic layer to black-box systems to add transparency (Guidotti et al., 2018; Chakraborty et al., 2017; Tjoa and Guan, 2015).

User-centred XAI, which is concerned with aspects related to user-interaction and experience (Ribera Turró and Lapedriza, 2019), is mainly focused on phases 2 and 3 and aims to integrate a user into the loop of an AI system's decision making as much as possible (Anjomshoae et al., 2019). Phase 2 deals with what is exactly to be provided to the end-user and how to present it, while phase 3 is concerned with the level of understanding that is achieved in an end-user with an explanation.

For these varying tasks identified within system-centred and user-centred XAI, it is useful to consider which argumentation theoretical framework can best provide the output that is most effective in a particular setting. In this paper, we briefly discuss the roles that some of the main argumentation theories can play for both system-centred and user-centred XAI approaches. Section 2 presents the role of Dung's theories and Walton's dialogue for achieving system-centred XAI, while Section 3 explores how Pragma-dialectics and Inference Anchoring Theory contribute towards user-centred XAI. Finally, Section 4 makes some final observations about the suitability of each theory to XAI.

2 System-centred XAI

Most of the literature on system-centred XAI does not differentiate between *interpretability* and *explainability* of learning models. Guidotti et al.

* Both M.D. and Q.S. contributed equally in the writing of this paper. Specifically, M.D. focused on Sections 3 and 4 and Q.S. focused on Sections 1 and 2.

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 44–49,
Dublin, Ireland, 18 December 2020.

(Guidotti et al., 2018) consider *explainability* as an interface between interpretable models and human users. They formalise four types of explanations for black boxes: (1) simulating them with an equivalent symbolic model that explains its working, (2) explaining only the black-box outcome rather than its working, (3) providing visual representation of the black-box mechanism for inspection, and (4) a transparent model that is fully explainable on its own without needing any supplementary explanations model. Rudin makes a distinction between *interpretable* ML and *explainable* ML (Rudin, 2019) where the latter involves the first three types of explanations as identified by Guidotti et al. while the former includes the last type. Based on this discussion, recent approaches to system-centred XAI can be classified into two main types: *interpretable* and *non-interpretable*. Interpretable black-box models can either be purely symbolic models or grey-box models, that is, those that generate intermediate symbols which can be leveraged for generating a trace of the reasoning process used by the model. Non-interpretable models will then refer to black-boxes for which only input and output are available. Thus, achieving explainability nails down to answering the question of how to generate or extract the symbols out of the black-boxes that make up the explanations. In the next two sections, we explore some preliminary ideas on how Abstract Argumentation Framework (AF) and dialogue theory can help us answer this question.

2.1 Abstract Argumentation Framework

An AF as defined by Dung (Dung, 1995) is a pair, $\langle A, R \rangle$, where A is a set of arguments and R is a binary relation on the set A which determines the attack relations between the arguments (Baroni and Giacomin, 2009). The arguments in an AF are atomic entities without any internal structure. This abstraction allows generalising the properties of the framework independently of the internal argument structure and possibly re-using the argumentation model across specific problems. AFs can be used as the formalism underpinning explanations for a black-box as we can see next.

For any black-box model, the data can be classified into three types: input, output and intermediate symbols which are generated during the learning process. Given such a black-box model, we consider different routes to XAI. A simple one would be to use a decision tree based approach as an initial step to build an AF. First, we apply a classification algorithm such as ID3 (Quinlan, 1986) over a table that contains the input (and possibly the intermediary data) as features of items and the output as their classes. The arguments of the AF could then be extracted from the decision tree. The labels (arguments) in A would be any subset of the nodes of the tree, including singletons. The label (argument) of the set of nodes in a path leading to a class C_i would attack the label representing the set of nodes of a path leading to a different class C_j. Other attack relations could be found, as well as other relationships for variants of Dungs model like Bipolar argumentation systems (Cayrol and Lagasquie-Schiex, 2005). For instance, labels representing the nodes in paths leading to the same class C_i support each other. Then explanations of an output (a class) can become the arguments (nodes and paths) in some preferred semantics over AF. This approach makes sense only if the input data is symbolic.

Figure 1 shows the schema of the data set for the classification algorithm.Each row in the table corresponds to a training example. Each column represents a feature label. *Input features* represent the input (independent) variables represented by i_{pn} where $p \in$ row number and $n \in$ column number. *Intermediary features* represent the intermediate symbols such as outputs of hidden layers generated from a black box model such as a neural network. These are represented by m_{pm} where $p \in$ row number as before and $m \in$ column number. *Output class* indicates the corresponding classification label for each row, represented by c_p

2.2 Dialogue Theory

Dialogue theory in argumentation can play a vital role in bridging the explanation gap between machine recommendations and human trust and understanding of these. During a dialogue, one party is typically seeking to persuade the other party to agree to some disputed conclusion. In contrast, while providing an explanation, one party is trying to provide some information to the other party in order to improve understanding of the already accepted conclusion (Walton, 2009). In this context, argumentation dialogues can be used to query black-box models on their intermediate symbols in order to generate more enriched explanation models. For example, consider a hypothetical decision system on the lines of COMPAS (Larson et al., 2016) which recommends parole or not for convicts on the basis of past parole violations and age. The

Input features				intermediary features				Output class
i_{11}	i_{12}	...	i_{1n}	m_{11}	m_{12}	...	m_{1m}	c_1
i_{21}	i_{22}	...	i_{2n}	m_{21}	m_{22}	...	m_{2m}	c_2
...	...	...	...	...	...	...	...	...
i_{p1}	i_{p2}	...	i_{pn}	m_{p1}	m_{p2}	...	m_{pm}	c_p

Figure 1: Input table for a classification algorithm. Sets of features become arguments in an AF.

system can be queried for an explanation of specific outcomes such as 'Why is this parole granted?' The system could use the features used for recommendation as justification such as 'Because there are no past parole violations'. In this case, the user was able to gain some information from the system.

Another scenario could be a case where the explanation model poses a question to the AI system regarding a feature which the decision system has not considered. For example, assuming that there is a query from the user to justify the outcome for the hypothetical parole system such as 'Is it because of my ethnicity?' In this case, ethnicity is not something the system has taken into account. So the system can try to find the symbols that can help it to determine the correlation and inform the user accordingly. In this way, the system is forced to look for more information resulting in not only a more enriched explanation model for the user but also more transparency for the system as it can cause hidden biases and correlations to be identified. Both these examples fall under the *Information Seeking Dialogue* type proposed by Walton where the dialogue goal is an information exchange. The argument generation approach from Section 2.1 can be combined with dialogue generation in the manner of Walton to explain black-box models as highlighted in this section.

3 User-centred XAI

User-centred XAI focuses on the way explanations generated by AI systems are communicated to non-expert and non-technical users, who often do not require a full understanding of the inner workings of the system they are interacting with. Instead, this type of user will be primarily interested in natural language explanations that are maximally understandable, that build trust and confidence in the system's recommendations, and that inform a user about how to alter the outcome of a decision (Hind, 2019). For example, when an AI system rejects a user's application for a bank loan, it should explain in natural language which variable (e.g. salary, or outstanding debt) is responsible for this output and what is needed in order to be eligible for a loan.

Providing explanations to non-expert users of AI systems is widely recognised as an essential component in XAI, but adapting these explanations to the particular needs of a user and the communicative setting in which they occur remains a challenging task (Anjomshoae et al., 2019). In order to endow AI systems with trustworthy and realistic interactive capabilities it is necessary to model the dialogical setting in which users interact with AI systems and to determine which types of communication and reasoning are most effective for informing users. The following two sections discuss the pragma-dialectical theory of argumentation and Inference Anchoring Theory, which are theoretical frameworks for modelling argumentation and reasoning in natural language.

3.1 The Pragma-dialectical Theory of Argumentation

The pragma-dialectical theory of argumentation (Van Eemeren and Grootendorst, 2004) is designed to allow for the analysis and evaluation of argumentation as it is actually used in communication. In pragma-dialectics, argumentation is considered as a complex and interlinked array of speech acts, which are directed towards fulfilling the ultimate goal of a critical discussion: the reasonable resolution of a conflict of opinion. Ideally, a discussion consists of four dialectical stages, which are (1) the confrontation stage, (2) the opening stage, (3) the argumentation stage and finally, (4) the concluding stage. In the confrontation stage, arguers establish they have a difference of opinion, which they may decide to attempt to resolve in the opening stage. The argumentation stage is dedicated to providing arguments in support of the standpoints proposed by the arguers and in the concluding stage the parties determine whether their difference of opinion has been resolved and in who's favour (Van Eemeren and Grootendorst, 2004, pp. 59-62).

In the context of user-centred XAI, this allows us to determine how the exchange of messages between a system and a user should be specified at different stages of communication, e.g. an explanation should be differently communicated depending on whether a message is provided in the confrontation stage or the argumentation stage.

The pragma-dialectical theory also stipulates ten rules for a critical discussion (Van Eemeren and Grootendorst, 2004, pp. 190-196), which represent the conditions arguers must uphold in order to ensure a reasonable discussion. Any violation of these critical discussion rules constitutes a hindrance towards the reasonable resolution of the conflict of opinion and is considered a fallacious argumentative move. These rules for a critical discussion reflect the normative element of the pragma-dialectical theory of argumentation and allow for an evaluation of the reasonableness of argumentation in actual language use. As such, the pragma-dialectical theory makes it possible to model the dialogical setting in which a user-AI interaction takes place and to establish whether the arguments that are used are fair and suited to the intended goals of an AI system's end user.

3.2 Inference Anchoring Theory

Inference Anchoring Theory (IAT) (Budzynska and Reed, 2011) is a theoretical framework for connecting the inferential structures that are present in argumentation with their associated illocutionary forces and dialogical processes. Consider the following example, taken from (Budzynska and Reed, 2011):

(1) a. Bob: *p is the case*
b. Wilma: *Why p?*
c. Bob: *q*.

Example (1) contains a dialogical structure that represents the order in which the propositions were uttered, which is governed by dialogical rules that stipulate how the participants in the dialogue may make communicative moves. This locutionary level (i.e. what is actually being said) of the dialogue and the transitions between the statements made are represented on the right-hand side of the diagram shown in Figure 2. Additionally, (1) can be viewed as containing a basic inferential structure, including a premise (p) and a conclusion (q). This propositional content and its logical structure are represented on the left-hand side. Central to IAT, the propositional content of a dialogue is 'anchored' in its respective locution or transitions through an illocutionary connection which represents the illocutionary force (Searle, 1969) that is exerted with a particular statement (e.g. asserting, arguing, or promising) and is represented in the middle of the diagram.

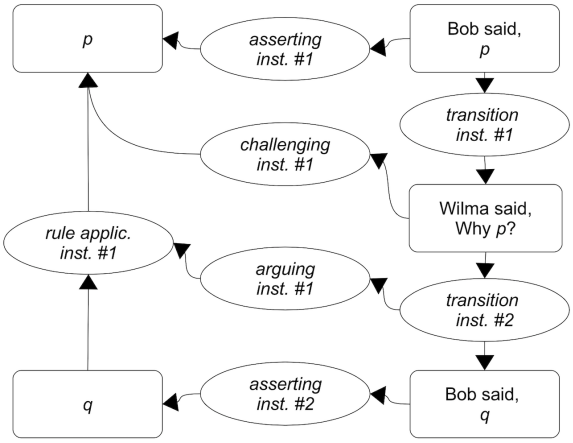

Figure 2: Interaction between argument and dialogue in IAT (Budzynska and Reed, 2011).

In summary, Inference Anchoring Theory allows to unpack four dimensions of explanations which can be then differently computed: it is possible to link (1) dialogical acts ("Bob said: *p is the case*") to (2) their propositional contents (p) through (3) an illocutionary connection that signifies the communicative intention of the speaker/user (*asserting instance #1*) linked to (4) ethotic conditions that allow us to express the credibility, trustworthiness, and character of a speaker (user modelling). This is particularly valuable for the task of user-centred XAI, since it enables the adaptation of argumentation and explanation to specific users.

4 Discussion and Future Work

In this paper, we have differentiated between system-centred and user-centred XAI, and discussed how four major argumentation theoretical frameworks can be applied to these challenges. Depending on the type of explanation required from an AI system, it is useful to consider the various theoretical tools that these approaches offer. Abstract Argumentation and dialogue theory excel in generating explanations of the inner workings of an AI system and modelling inter-system interaction. Pragma-dialectics and Inference Anchoring Theory are especially suited towards modelling the dialogical setting of human-AI interaction and identifying which type of reasoning is most effective there.

Future work on system-centred XAI could explore how Abstract Argumentation Framework and dialogue theory can be used in a multi-agent recommender system. In this case, the goal is to achieve explainability for the joint recommendation made by multiple systems after consensus. However, in order to achieve consensus, we need dialogue between the different systems. In this context, we can explore using Abstract Argumentation Framework for justifying the recommendation and dialogue theory for achieving consensus on the recommendation itself.

For user-centred XAI, we propose to investigate how pragma-dialectics and Inference Anchoring Theory can be applied for modelling users in social media. To this end, Natural Language Processing techniques such as argument mining can help create an image of a user's linguistic profile, which provides insight into their communicative behaviour and reasoning patterns (i.e. argumentation schemes). In turn, these argumentation schemes can form a blueprint for the generation of arguments and explanations that are tailored to a specific communicative situation and a particular user. In that capacity, argumentation schemes carry substantial value for tasks in explainable AI related to language generation, inter-agent communication, and personalising AI systems to end users.

To conclude, in order to further improve our understanding of, and our interaction with AI systems, we believe it is fruitful to build on existing argumentation theoretical frameworks in various ways towards more robust and accurate methods for eXplainable Artificial Intelligence.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621.

References

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1078–1088, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Pietro Baroni and Massimiliano Giacomin. 2009. Semantics of abstract argument systems. In *Argumentation in Artificial Intelligence*, pages 25–44. Springer US.

Katarzyna Budzynska and C. Reed. 2011. Whence inference. Technical report, University of Dundee.

C. Cayrol and M. C. Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3571 LNAI, pages 378–389. Springer Verlag.

S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram. 2017. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6.

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(2):321–357.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5).

Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):16–19.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 201 6. How We Analyzed the COMPAS Recidivism Algorithm — ProPublica.

Mark A. Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. 2018. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *Engineering Psychology and Cognitive Ergonomics*, pages 204–214, Cham. Springer International Publishing.

J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

Mireia Ribera Turró and Agata Lapedriza. 2019. Can we do better explanations? A proposal of User-Centered Explainable AI.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Erico Tjoa and Cuntai Guan. 2015. A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI. Technical Report 8.

Frans H. Van Eemeren and R. Grootendorst. 2004. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Douglas Walton. 2009. Argumentation Theory: A Very Short Introduction. In *Argumentation in Artificial Intelligence*, pages 1–22. Springer US, Boston, MA.

Toward Natural Language Mitigation Strategies for Cognitive Biases in Recommender Systems

Alisa Rieger
TU Delft
Van Mourik Broekmanweg 6
2628 CD Delft
a.rieger@tudelft.nl

Mariët Theune
University of Twente
Drienerlolaan 5
7522 NB Enschede
m.theune@utwente.nl

Nava Tintarev
TU Delft
Van Mourik Broekmanweg 6
2628 CD Delft
n.tintarev@tudelft.nl

Abstract

Cognitive biases in the context of consuming online information filtered by recommender systems may lead to sub-optimal choices. One approach to mitigate such biases is through interface and interaction design. This survey reviews studies focused on cognitive bias mitigation of recommender system users during two processes: 1) item selection and 2) preference elicitation. It highlights a number of promising directions for Natural Language Generation research for mitigating cognitive bias including: the need for personalization, as well as for transparency and control.

1 Introduction

Decision-making at an individual, business, and societal levels is influenced by online news and social media. Filtering and ranking algorithms such as recommender systems are used to support these decisions. Further, individual cognitive selection strategies and homogeneous networks can amplify bias in customized recommendations, and influence which information we are exposed to (Bakshy et al., 2015; Baeza-Yates, 2018).

Biased exposure to online information is known to accelerate extremism and the spread of misinformation (Hills, 2019). Ultimately, these undesirable negative consequences of information filtering diminish the quality of public discourse and thus can pose a threat to democracy (Bozdag and van den Hoven, 2015).

One strategy for bias mitigation would be to raise users' awareness of filtering mechanisms and potential cognitive biases. Approaches going one step further than creating awareness, actively nudge users in a direction of less biased information selection and diversification. Explanations and nudges for mostly non-expert users of recommender systems in the domains of news and social media have to be designed in a way that they are understood intuitively, e.g., using natural language (Liao et al., 2020).

To our knowledge, no previous work has summarized cognitive bias mitigation in the context of recommender systems. In this paper, we aim to identify research gaps and opportunities to improve natural language explanation interfaces that mitigate cognitive biases. We do this by providing an overview of approaches to mitigate cognitive bias of recommender system users in the domains of news and social media. We review the literature in the field and summarize ways of measuring bias and mitigation approaches for different biases in different contexts. We also consider how these occur at different stages of the recommendation process. In sum, we address the following research questions (RQs):

1. For which types of cognitive biases occurring among users of recommender systems exist validated mitigation approaches?
2. What are effective approaches to *measure* different types of bias?
3. What are effective approaches to *mitigate* different types of bias?
4. How are the mitigation approaches *evaluated*?

In the next section, we introduce the method used in our literature review. Then, in Section 3, we analyze the resulting papers and identify commonalities. We see that human bias mitigation using natural language generation in recommender systems is still under-explored despite explanations being successfully applied in the fields of persuasive technology and argumentation (Dragoni et al., 2020; Guerini et al., 2011). So, in Section 4 we take a constructive approach and discuss promising directions for natural language generation (NLG) research, before concluding in Section 5.

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NLXAI 2020), pages 50–54,
Dublin, Ireland, 18 December 2020.

2 Methodology

To find relevant literature for this survey, we defined inclusion criteria as a search string which we ran through the databases Springerlink (http://link.springer.com) and ACM digital library (https://dl.acm.org) in July 2020. These two databases are established and comprehensive databases in the field of computer science, and support complex search strings. The search results were filtered by scanning Title, Abstract, and Discussion.

Inclusion criteria: Our search string covers four main concepts: **(1)** bias-related; **(2)** target-system-related; **(3)**; domain-related; **(4)** mitigation-related. The terms used for each concept are: **(1)** *("cognitive bias" OR "human bias" OR "confirmation bias" OR "availability bias" OR "backfire effect" OR "homophily" OR "affinity bias" OR "decoy effect" OR "selective exposure" OR "false consensus effect" OR "saliency bias") AND* **(2)** *("recommender" OR "recommendation") AND* **(3)** *("news" OR "social media" OR "search" OR "information seeking") AND* **(4)** *("mitigat*" OR "debiasing" OR "reduce" OR "explainable artificial intelligence" OR "XAI" OR "intelligent user interface" OR "IUI" OR "natural language").* This search resulted in 257 hits.

Exclusion criteria: Papers are excluded if they do not: **a)** focus on recommender systems in the domains of news, social media, or search (40 excluded); **b)** do not propose a mitigation approach for human bias (137); **c)** do not present a user study (66); **d)** do not include measures of bias (5); **e)** we have no access to the full paper (5). These criteria lead to the exclusion of 253 papers, resulting in the four papers discussed in the remained of this paper (see Table 1). We observe that these papers do not cover linguistic solutions, but will later see that they still highlight promising areas for research in NLG.

3 Analysis

In this section we analyze and compare the four resulting papers based on five aspects which were chosen to answer the research questions: **(RQ1)** *Objective*: context and objective of the paper and *Bias*: type of cognitive bias investigated; **(RQ2)** *Measure*: approach for measuring bias; **(RQ3)** *Mitigation*: approach of bias mitigation; and **(RQ4)** *Evaluation*: evaluation of the mitigation approach and moderating factors.

(RQ1) *Objective and Bias*: To encourage diverse information and common ground seeking, Liao and Fu (2014) investigated the mitigation of selective exposure or the confirmation bias, which is the tendency to search for and select information which confirms previous beliefs and values, in online discussion forums. Graells-Garrido et al. (2016) researched the mitigation of confirmation bias and homophily, the tendency to have and build ties to similar individuals to oneself, with the intention to connect users with different opinions in social networks. Tsai and Brusilovsky (2017) studied the mitigation of homophily and position bias, occurring if the position influences the perceived value or utility of an item, in the context of a tool for conference attendees to connect to diverse scientists. Pommeranz et al. (2012) intended to design user interfaces for unbiased preference elicitation, which are needed for accurate recommendations. Preference elicitation describes the process of collecting user data to build an accurate user-model, based on which items are recommended. Thus, Pommeranz et al. (2012) investigate bias mitigation at an earlier stage in the recommendation process, than the other three reviewed studies. The authors list a number of possible biases that can occur during the stage of preference elicitation (but do not measure them): *framing* – presentation with positive or negative connotations influence the perceived value or utility of an item, *anchoring* – value of an initially encountered item influences the perceived value of a subsequently encountered item, and *loss aversion* – tendency to prefer avoiding losses to obtaining gains with the same value.

(RQ2) *Measure*: To measure bias, all of the studies compared the effect of an intervention with a baseline system on a set of metrics. For the three studies researching confirmation bias and homophily during item selection, the diversity of item selection or the degree of exploration of items was compared to the baseline (without bias mitigation) (see Liao and Fu, 2014; Graells-Garrido et al., 2016; Tsai and Brusilovsky, 2017). Diversity and degree of exploration were calculated on basis of the users' clicking behavior and attributed values for each item, reflecting the aspects of interest in the study (e.g., position - pro/con, similarity of profile - high/low,..). For framing, anchoring, and loss aversion during preference elicitation, a quality score was calculated for each tested preference elicitation method. A high level of agreement between the system's outcome preference model and the user-generated list of preferences resulted in a

	Bias	Objective	Mitigation
Liao and Fu, 2014	confirmation bias	viewpoint diversification of users in forum for political discussions	*Visual barplot*: indication of source position valence and magnitude to reduce the demand of cognitive resources
Graells-Garrido et al., 2016	confirmation bias and homophily	connecting users with diverse opinions in social networks	*Visual data portraits and clustering*: indication of own interests and opinions as data portrait to explain recommendations, and display of users with shared latent topics in interactive clusters to facilitate exploration
Tsai and Brusilovsky, 2017	homophily and position bias	help conference attendees to connect to diverse scientists via a social network	*Multidimensional visual scatterplot*: display of scientists' accademic and social similarity and highlights potential matches through color-coding
Pommeranz et al., 2012	framing, anchoring, loss aversion	designing user-centered interfaces for unbiased preference elicitation	*Multiple visual interface proposals*: virtual agent with thought bubble, outcome view (explore link between interests, preferences and outcomes), interest profiling, affective feedback,..

Table 1: Examined Bias, Objective, and Mitigation approach per paper

high quality score (see Pommeranz et al., 2012).

(RQ3) *Mitigation*: Liao and Fu (2014) displayed posts in the online forum in combination with a visual barplot which indicated position valence (pro/con) and magnitude (moderate/extreme) of the posts' authors to mitigate confirmation bias. The authors argue that freeing up cognitive resources can increase users capacity to assess viewpoint challenging information. They aimed to reduce the demand on cognitive resources by pre-evaluating and marking the author's position, with the intention that this would increase users' capacity to process information relating to the post's *content*.

Further, the explicit indication of author position information aimed at encouraging attention to diverse viewpoints and motivating users to select attitude-challenging information. Graells-Garrido et al. (2016) recommended diverse profiles with shared latent topics and displayed visualizations of the user's own data portrait in the form of word-clouds with interests and opinions to explain the given profile recommendations and mitigate confirmation bias and homophily. Profile recommendations were presented in the form of visual clusters of accounts with shared latent intermediary topics, from which the user could select accounts for exploration. This approach aimed to overcome cognitive dissonance produced by direct approaches of exposure to challenging information. The aim was to provide context to a given recommendations, both in form of the user's own data profile and the basis of a shared intermediary topic, to give the new connection a chance. Another approach to mitigate homophily in addition to position biases was chosen by Tsai and Brusilovsky (2017), who presented scientists as points in a two-dimensional scatterplot. The position of a point was calculated by social (co-authorship) and academic (publication content) feature similarity (0 - 100 %) between user and scholar. Meaningful feature combinations, defined by higher degrees of feature similarities, were highlighted through color-coding. This approach aimed to enable the presentation of more than one recommendation aspect, to guide conference attendee's attention to areas of scientists with meaningful feature combinations, and overall, to promote diversity of profile exploration. Pommeranz et al. (2012) propose input methods and interfaces for preference elicitation which result in equal mental preference model and system preference representation to achieve a mitigation of framing, anchoring and loss aversion biases. They investigated different methods of preference elicitation, such as rating with a nine point likert scale (like to dislike), ordering, navigational (receiving immediate feedback after changing preference for one item), and affective rating.

In summary, the mitigation approaches of confirmation bias and homophily use the visual display of information to increase users' awareness for item-features of interest (e.g., position valence, similarity,..) and to encourage and facilitate the intuitive exploration of diverse items. Approaches include multidimensional feature representation plots, and additional highlighting in form of color-coding or clustering of meaningful feature combinations. Two studies aim to enable users to understand contingencies between preferences, item selections and recommendation outcome and thus to a certain degree explaining recommendations. They do this by visually displaying the system's user model in form of a word cloud or an interest profile, preference summary, value chart or outcome view.

(RQ4) *Evaluation*: On their attempt to mitigate

confirmation bias, Liao and Fu (2014) measured the potentially moderating factor of accuracy motive (motivation to accurately learn about a subject) of the users before exposure to the online forum. Results of the user study show that accuracy motive and position magnitude (moderate/extreme) of authors were functioning as moderating factors by influencing the effectiveness of bias mitigation. The authors conclude that interfaces should be individually adapted for users with varying levels of accuracy motive and that authors with moderate opinion could function as bridges between users with different opinions. Graells-Garrido et al. (2016)'s clustered visualization of recommendations, aiming to mitigate confirmation bias and homophily, was found to be effective in increasing users' exploration behavior (users clicked on more diverse items). The proposed recommendation algorithm based on shared latent topics, however, was not effective in increasing exploration behavior. The results show that political involvement of the users was functioning as a moderating factor, influencing the effectiveness of bias mitigation. Thus, Graells-Garrido et al. (2016) conclude that no one-size-fits-all solution exists, but that indirect approaches of transparent recommendations and user profiles rather than directly exposing users to opposing information should be considered for bias mitigation. Results of Tsai and Brusilovsky (2017)'s study on mitigating homophily and position biases show, that the exploration patterns were more diverse in the experimental conditions of presenting scientists in a multi-dimensional scatterplot compared to a baseline of displaying them in a ranked list. However, in a post-experimental questionnaire users reported a higher intent to reuse the ranked list than the multi-dimensional scatterplot. The authors conclude that diversity-oriented interfaces on the one hand can encourage the exploration of more diverse recommendations, but on the other hand can also impair intent to reuse the system and thus should be designed with care. The results of Pommeranz et al. (2012)'s user study on mitigating framing, anchoring and loss aversion during preference elicitation, show cognitively less demanding rating tasks were liked most and resulted in highest quality outcome lists. They conclude, that the interface design needs to adapt to individual differences in terms of user preferences. The authors highlighted the importance of transparency and control on the grounds that users found it very useful to be allowed to investigate the links between their interests, preferences and recommendation outcomes.

In summary, multiple studies highlight that no one-size-fits-all mitigation approach exists due to moderating user-related factors, such as the accuracy motive, diversity seeking or challenge averseness, motivation, political involvement and opinion. Thus the authors emphasize that interfaces should thus be designed to be personalizable. In addition, the need for transparent and interactive interface designs which allow control of user-profile and recommendations was highlighted.

4 Discussion

In this paper, we reviewed interface-based approaches for the mitigation of confirmation bias, homophily, position bias, framing, anchoring, and loss aversion **(RQ1)**. To measure bias, the studies compared the effect of an intervention with a baseline system on a set of metrics **(RQ2)**. The reviewed studies applied interactive multidimensional visualizations, rearranging, sorting, and highlighting through color-coding and size to increase users' awareness for diverse features, to facilitate and increase exploration of recommended items, and to align the system's user model with the user's mental preference model **(RQ3)**. During the evaluation of the approaches **(RQ4)**, multiple user-related factors that *moderated* the effectiveness of the reviewed mitigation approaches were identified. Consequently, the studies highlighted the need for personalized interfaces that can adapt to these factors. They include users' accuracy motive, motivation, political involvement, and prior opinions on recommended items or topics, all measured with tailor-made questionnaires or inferred from the user's behavior. Overall, transparency, control, as well as immediate feedback were found to enhance the users' understanding and to mitigate cognitive bias.

While the surveyed methods are within graphical interfaces, they help to uncover research questions for future studies in all interactive interfaces, also for *natural language-based* mitigation strategies:

1. Which approaches of interactive natural language bias mitigation approaches are most effective?
2. In which form and to which extent should transparency and control be given to the users?
3. What are user-related moderating factors and how could they be measured?

4. How could an interface personalization according to these user-related factors look like?

Our literature review also suggests that bias mitigation strategies using natural language could be used at different stages of interaction: **a)** conversational preference elicitation, **b)** pre-evaluation and explanation of recommended items, or **c)** to motivate behavior modifications for bias mitigation. Such interactions could promote the users' understanding of their profiles and the functioning of the system. Using NLG to increase user-control on the user-profile, algorithmic parameters, and the recommendation outcomes (Jin et al., 2020), appears to be a promising way to mitigate cognitive biases.

5 Conclusion

The analysed studies demonstrate effective approaches of implementing and evaluating interface-based cognitive bias mitigation for recommender system users. On this basis, we suggest promising areas for future research for bias mitigation using interactive NLG: personalization of explanations, and more immediate transparency and control.

Acknowledgments

This work has received funding from the *European Union's Horizon 2020* research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621.

References

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, 17(4):249–265.

Mauro Dragoni, Ivan Donadello, and Claudio Eccher. 2020. Explainable ai meets persuasiveness: Translating reasoning results into behavioral change advice. *Artificial Intelligence in Medicine*, page 101840.

Eduardo Graells-Garrido, Mounia Lalmas, and Ricardo Baeza-Yates. 2016. Data portraits and intermediary topics: Encouraging exploration of politically diverse profiles. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 228–240.

Marco Guerini, Oliviero Stock, Massimo Zancanaro, Daniel J O'Keefe, Irene Mazzotta, Fiorella de Rosis, Isabella Poggi, Meiyii Y Lim, and Ruth Aylett. 2011. Approaches to verbal persuasion in intelligent user interfaces. In *Emotion-Oriented Systems*, pages 559–584. Springer.

Thomas T Hills. 2019. The dark side of information proliferation. *Perspectives on Psychological Science*, 14(3):323–330.

Yucheng Jin, Nava Tintarev, Nyi Nyi Htun, and Katrien Verbert. 2020. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction*, 30(2):199–249.

Q Vera Liao and Wai-Tat Fu. 2014. Can you hear me now? Mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 184–196.

Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Alina Pommeranz, Joost Broekens, Pascal Wiggers, Willem-Paul Brinkman, and Catholijn M Jonker. 2012. Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction*, 22(4-5):357–397.

Chun-Hua Tsai and Peter Brusilovsky. 2017. Leveraging interfaces to improve recommendation diversity. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 65–70.

When to explain: Identifying explanation triggers in human-agent interaction

Lea Krause and **Piek Vossen**
Computational Lexicology & Terminology Lab (CLTL)
Vrije Universiteit Amsterdam
{l.krause,p.t.j.m.vossen}@vu.nl

Abstract

With more agents deployed than ever, users need to be able to interact and cooperate with them in an effective and comfortable manner. Explanations have been shown to increase the understanding and trust of a user in human-agent interaction. There have been numerous studies investigating this effect, but they rely on the user explicitly requesting an explanation. We propose a first overview of when an explanation should be triggered and show that there are many instances that would be missed if the agent solely relies on direct questions. For this, we differentiate between direct triggers such as commands or questions and introduce indirect triggers like confusion or uncertainty detection.

1 Introduction

The introduction of artificial agents into our daily lives means that an increasing number of lay users interact with them, often even collaborating. This has major societal implications since they are used in domains ranging from healthcare over finance to the military. As a result, special care must be taken to ensure that users understand agents' decisions, can effectively collaborate with them, and even hold them accountable if necessary.

While research emphasising the need for explanations is not new (Buchanan and Shortliffe, 1984), interest has picked up over the past few years (Anjomshoae et al., 2019). Recent advances in artificial intelligence and machine learning have led to a rapid increase in quality of artificial agents. Since most state-of-the-art models are black boxes, it is often not clear to the end-user why the agent made certain decisions. Trust, however, relies on understanding the decision-making of the agent (Lee and Moray, 1992) and trust is a prerequisite for successful collaboration and use. Explanations have been shown to increase the understanding of the agent in human-agent teams (Dzindolet et al., 2003; Wang et al., 2016) and thus increase trust. Within human-human interaction, people resolve conflicts or uncertainties by explaining the reasoning behind their arguments or decisions. Users have a tendency to anthropomorphise agents (Lemaignan et al., 2014) and expect them to behave human-like; thus, they expect them to give explanations for their decisions and actions.

Most work assumes that the user directly asks for an explanation (Sridharan and Meadows, 2019; Ray et al., 2019). We claim that there are many situations where explanations are needed, even if not explicitly requested by the user. In our work, we aim to provide an overview of direct as well as indirect explanation triggers. This overview will be the basis of designing future system experiments and evaluation metrics that target explanations to those needs.

While our primary goal is to investigate this in the context of human-robot interaction, we believe that the impact of these findings is not limited solely to this domain.

2 Related work

In this section, we review recent papers covering explainability, explanations and explanations specifically for human-agent interaction. As our focus lies on human-agent interaction we will mostly refer the reader to survey papers for the parts on explainability and explanations as they give a much more in-depth overview than what would be possible within this space.

2.1 Explainability

Recent years have seen the fundamental expansion of machine learning techniques starting within academia and spreading across industries. While these black-box models bring state-of-the-art results across domains, they are criticised for their

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NLXAI 2020), pages 55–60,
Dublin, Ireland, 18 December 2020.

biases and lack of transparency. The rapid rise of black-box models has resulted in a simultaneous surge of explainability methods. These methods aim to increase the transparency of the models and to make them explainable to humans. Going as far as to include "the right to explanation" in the European Union General Data Protection Regulation (GDPR) (noa, 2016). Adadi and Berrada (2018) have broken the need for explainable artificial intelligence down into four reasons: explain to justify, explain to control, explain to improve and explain to discover. The last two especially show that explainability does not need to slow a model down, but can instead further its development and share new discoveries it has made.

Although there has been a large number of publications in explainable artificial intelligence in recent years, no common taxonomy or agreed meaning has emerged. Two recent in depth proposals were done by Lipton (2016) and Sokol and Flach (2020). The latter propose a fact sheet detailing five dimensions to guide the development of future explainability approaches: 1. functional requirements, 2. operational requirements, 3. usability criteria 4. security, privacy and any vulnerabilities, 5. validation. Their approach is one of the few taking results from other disciplines, such as sociology and psychology, into account, which have been studying explainability and explanations much longer than artificial intelligence.

This lack of consideration of input from other disciplines is the topic of a thorough critique of the current state of explainable artificial intelligence by Mittelstadt et al. (2019). They examine the discrepancy between what designers and end-users want from explanations and come to the conclusion that explanations as they currently exist in artificial intelligence fail in providing adequate explanations to those affected by the results of the machine learning algorithms. Their recommendations to resolve this discrepancy are based on Miller (2019) whose findings we will discuss in the next paragraph.

2.2 Explanations

Explanations differ from general explainability in that they focus only on explaining a single prediction instance of a model or in our case, agent. The most extensive review of explanations within A.I. in recent years has been done by Miller (2019). He reviews existing research on explanations from social sciences, philosophy, psychology and cognitive science, and connects it to the current discourse in explainable artificial intelligence. His main conclusion is that explanations need to be contextualised instead of just stating a causal relation. He breaks this down into four findings:

1. An explanation should be contrastive, they are an answer to the question *Why did **A** happen instead of **B**?*

2. The selection of an explanation is biased; selected causes are chosen to fit the explanation

3. A probability alone does not make an explanation.

4. An explanation is part of a social interaction, related to the mental states of the participants of the conversation.

2.3 Human-agent interaction

Explanations for human-agent interaction often form a challenging task. They have to be generated in different circumstances with somewhat unpredictable input (unpredictable humans) and most people the agent will interact with are not experts, therefore the explanations have to be understandable for a lay-person.

Anjomshoae et al. (2019) have conducted a large-scale literature review on current literature (after 2008) on explainable agents and robots. Similarly to the field of explainability in general, they have found a rapid increase in works published since 2016. The similarities continue, as only 37% of the papers made any reference to the theoretical background of explanations. The main direction found to be relevant for future work is the communication of the explanations.

Rosenfeld and Richardson (2019) propose a taxonomy for explainability in human-agent systems in which they cover the questions of: *Why* is there a need for explainability?, *Who* is the target audience?, *What* kind of explanation should be generated? *When* should the explanation be presented to the human? and lastly *How* can the explanations be evaluated?

Another overview from a different angle was done by Sridharan and Meadows (2019). While they as well give a framework for explanations in human-robot collaboration, their main contribution is their investigation of combining knowledge representation, reasoning, and learning to generate interactive explanations.

Several other studies have investigated the effectiveness of explanations for task-oriented human-agent teams and reported an increased success rate and self-reported trust in the agent (Ray et al., 2019; Chakraborti et al., 2019; Gong and Zhang, 2018; Wang et al., 2016)

Recently, post-hoc explanations have been accused of fairwashing (Aïvodji et al., 2019) and Rudin (2019) specifically called for researchers to focus on completely interpretable models if it is a high stakes decision. Agents can be deployed in many circumstances, also high stake ones. We agree that only post-hoc explanations of blackbox models are not enough under these circumstances, but we believe that explanations are nevertheless important in the case of human-agent interaction as they fulfil a communicative function as well as an informative one.

3 Triggers

All the work on explanations for human-robot agents mentioned before makes the assumption that the user is explicitly going to ask for an explanation and to the best of our knowledge, the question when during communication an explanation is actually needed remains unanswered. Rosenfeld and Richardson (2019) pose the question in their overview paper on explainability in human-agent systems, but only consider the task of the agent-system, not the communicative aspect or any flexible trigger detection that takes the users current state into account. We argue that it is vital to fill this gap in order to make use of the full potential of explanations for human-agent interaction, as there are many situations in which an explanation is needed, even if not explicitly requested by the user. We therefore provide a first overview of possible direct and indirect triggers.

When users interact with explainable agents, the agent constantly has to evaluate whether it has to inform the user about its decisions. To do this efficiently it needs clear triggers when to explain.

3.1 Direct triggers

The most obvious triggers of explanations are explicitly expressed **commands or questions**. According to Miller (2019), an explanation is inherently an answer to a why-question. There are different underlying causes for such an explicit question. As described earlier, *trust* plays a significant role in human-agent interaction. One of the principal reasons for the user to demand an explanation is thus when they mistrust the decision of the agent and need clarification. Secondly, the user could be uncertain whether they have understood the agent correctly and seek an explanation to resolve this *uncertainty*. One step further is the occurrence of a *knowledge gap*. Here, the user might be completely unfamiliar with the topic of a decision. This case is also a critical one, as the user otherwise could not judge whether the decision is correct. Consequently, the reliability of the explanation is likewise of utmost importance. Lastly, it could simply be *curiosity*, due to interacting with a new agent or a topic that sparked the interest of the user. These examples also show that the agent should tailor its explanations to the underlying trigger.

Direct triggers	**Indirect triggers**
Command / Question	Confusion detection
Urgency	Agent uncertainty
	Conflicting mental states
	Conflict of interest
	Lack of trust

Table 1: Overview of direct and indirect triggers of explanations

We argue that additionally there are cases where an inherent **urgency** to inform is inherent to the topic without the occurrence of a question or uncertainty. This case is particularly relevant in the context of agents. The agent might observe something the human has overlooked. A situation like this would rely heavily on the agent's reasoning capabilities. It needs to analyse the situational urgency, the potential impact, and react instantaneously. Common use cases for this are agents deployed in elderly homes.

3.2 Indirect triggers

The often multi-modal nature of agents gives the opportunity to detect the need for an explanation not solely by relying on explicit commands. Indirect triggers largely depend on signal interpretation and belief detection. An example from educational computing is confusion detection (Arguel et al., 2017; Bosch et al., 2014). Detecting confusion is an especially fitting case for explainable agents, as they can also use it to detect whether an explanation was successful. Firstly, the agent can use **visual cues**, here we draw from Arguel et al. (2017) findings for detecting confusion in digital learning

environments. This can be *eye-tracking*, where the user's gaze is captured. Direction and duration of the user's eye movement can indicate their focus of attention as well as their emotional status. Eye-tracking is not deemed suitable for online learners, as to not add extra equipment. Agents, however, are often equipped with high-resolution cameras and object recognition, making them suitable for this type of detection. Another visual cue are *facial expressions*. Facial expressions have long been used in affect detection. Lowering the eyebrows paired with tightening the eyelids are indicators of confusion (D'Mello et al., 2009). *Body posture and movement* are further indicators. These can include shoulder position, hand placement and movements like head-scratching.

A second possible modality are **audio cues**. In *prosody*, rising intonation can indicate uncertainty and is more often paired with a wrong answer to a question than falling intonation (Brennan and Williams, 1995). *Speech disfluency*, like filler words such as "huh", "uh" or "um", occur more often if the speaker is uncertain or is presented with a choice. There is even a hierarchy as "um" marks a a greater uncertainty than "uh" (Brennan and Williams, 1995; Corley and Stewart, 2008).

An important step towards transparency is, that if the robot detects an **uncertainty**, be it from an unclear signal or the occurrence of multiple equally likely solutions, it gives an explanation why the decision might not be trustworthy.

The following triggers prompt explanations used for **reconciliation** between the agent and the user.

A more abstract trigger for an explanation can be found within **theory of mind** (Shvo et al., 2020; Miller, 2019). If the agent detects conflicting beliefs or mental states between itself and the user, it can take the user's beliefs into account and try to resolve them. These conflicting beliefs can be many fold, for one it can be distinguished between a misunderstanding and a misconception (McRoy and Hirst, 1995). A *misunderstanding* occurs when one side does not succeed in conveying the beliefs that they wanted to convey to their conversational partner. This is for example the case if a student misunderstands the question on an exam (Olde Bekkink et al., 2016). *Misconceptions* on the other hand are related to factual states of the world. Here the user could have an incorrect belief about what is a case in the world or what could be a case in the world (Webber and Mays, 1983), for instance believing that Northern Ireland is part of Great Britain.

The last potential trigger is a **conflict of interest**. In this case, there is a full understanding between the agent and the human, but a disagreement about the planning or the method to reach the goal. The agent needs to explain itself to either reach an agreement or for the user to be able to make an informed choice to disregard the agent's suggestion.

While we have described the triggers as separate entities, users will benefit most, if all of the signals are processed simultaneously by the agent.

4 Conclusion

We have shown the need for detecting triggers of explanations and given a first classification of possible internal and external triggers. Next steps will be to implement this classification system into an agent. Further work will then correlate the triggers to specific types of explanations and their generation.

5 Acknowledgments

This research was funded by the Vrije Universiteit Amsterdam, the Netherlands Organisation for Scientific Research via the Spinoza grant awarded to Piek Vossen and the Hybrid Intelligence Centre via the Zwaartekracht grant from the Dutch Ministry of Education, Culture and Science.

References

2016. General Data Protection Regulation (GDPR) – Official Legal Text.

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

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems.

Amaël Arguel, Lori Lockyer, Ottmar V. Lipp, Jason M. Lodge, and Gregor Kennedy. 2017. Inside Out: Detecting Learners' Confusion to Improve Interactive Digital Learning Environments. *Journal of Educational Computing Research*, 55(4):526–551. Publisher: SAGE Publications Inc.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp.

2019. Fairwashing: the risk of rationalization. In *Proceedings of the 36 th International Conference on Machine Learning,*, page 10, Long Beach, California.

Nigel Bosch, Yuxuan Chen, and Sidney D'Mello. 2014. It's Written on Your Face: Detecting Affective States from Facial Expressions while Learning Computer Programming. In *Intelligent Tutoring Systems*, pages 39–44, Cham. Springer International Publishing.

Susan E Brennan and Maurice Williams. 1995. The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language*, 34(3):383–398. Publisher: Elsevier.

Bruce G Buchanan and Edward H Shortliffe. 1984. Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley series in artificial intelligence). Publisher: Addison-Wesley Longman Publishing Co., Inc.

Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. 2019. Plan Explanations as Model Reconciliation – An Empirical Study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 258–266. ISSN: 2167-2148.

Martin Corley and Oliver W. Stewart. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

Sidney D'Mello, Scotty Craig, and Arthur Graesser. 2009. Multi-method assessment of affective experience and expression during deep learning. *International Journal of Learning Technology*, 4(3-4):165–187.

Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718.

Ze Gong and Yu Zhang. 2018. Behavior Explanation as Intention Signaling in Human-Robot Teaming. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1005–1011. ISSN: 1944-9437.

John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270. Publisher: Taylor & Francis.

Séverin Lemaignan, Julia Fink, and Pierre Dillenbourg. 2014. The Dynamics of Anthropomorphism in Robotics. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 226–227. ISSN: 2167-2121.

Zachary C. Lipton. 2016. The Mythos of Model Interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning*, New York, NY, USA.

Susan W. McRoy and Graeme Hirst. 1995. The Repair of Speech Act Misunderstandings by Abductive Inference. *Computational Linguistics*, 21(4):435–478.

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

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 279–288. ArXiv: 1811.01439.

Marleen Olde Bekkink, A. R. T. Rogier Donders, Jan G. Kooloos, Rob M. W. de Waal, and Dirk J. Ruiter. 2016. Uncovering students' misconceptions by assessment of their written questions. *BMC Medical Education*, 16(1):221.

Arijit Ray, Yi Yao, Rakesh Kumar, Ajay Divakaran, and Giedrius Burachas. 2019. Can you explain that? Lucid explanations help human-AI collaborative image retrieval. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 153–161. Issue: 1.

Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6):673–705.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. Publisher: Nature Publishing Group.

Maayan Shvo, Toryn Q. Klassen, and Sheila A. McIlraith. 2020. Towards the Role of Theory of Mind in Explanation. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 75–93, Cham. Springer International Publishing.

Kacper Sokol and Peter Flach. 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67. ArXiv: 1912.05100.

Mohan Sridharan and Ben Meadows. 2019. Towards a Theory of Explanations for Human–Robot Collaboration. *KI - Künstliche Intelligenz*, 33(4):331–342.

Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 109–116. ISSN: 2167-2148.

Bonnie Lynn Webber and Eric Mays. 1983. Varieties of user misconceptions: detection and correction. In *Proceedings of the Eighth international joint conference on Artificial intelligence - Volume 2*, IJCAI'83, pages 650–652, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Learning from Explanations and Demonstrations: A Pilot Study

Silvia Tulli
INESC-ID and IST, Portugal
silvia.tulli@gaips.inesc-id.pt

Sebastian Wallkötter
Uppsala University, Sweden
sebastian.wallkotter@it.uu.se

Ana Paiva
INESC-ID and IST, Portugal
ana.paiva@inesc-id.pt

Francisco S. Melo
INESC-ID and IST, Portugal
fmelo@inesc-id.pt

Mohamed Chetouani
ISIR and SU, France
mohamed.chetouani@sorbonne-universite.fr

Abstract

We discuss the relationship between explainability and knowledge transfer in reinforcement learning. We argue that explainability methods, in particular methods that use counterfactuals, might help increasing sample efficiency. For this, we present a computational approach to optimize the learner's performance using explanations of another agent and discuss our results in light of effective natural language explanations for both agents and humans.

1 Introduction

The process of gaining knowledge from the interaction between individuals needs to allow a two-way flow of information, i.e., reciprocally active communication. During this process explainability is key to enabling a shared communication protocol for effective information transfer. To build explainable systems, a large portion of existing research uses various kinds of natural language technologies, e.g., text-to-speech mechanisms, or string visualizations. However, to the best of our knowledge, few works in the existing literature specifically address how the features of explanations influence the dynamics of agents learning within an interactive scenarios.

Interactive learning scenarios are a much less common but similarly interesting context to study explainability. Explanations can contribute in defining the role of each agent involved in the interaction or guide an agent's exploration to relevant parts of the learning task. Here, some of the known benefits of explanability (e.g., increased trust, causality, transferability, informativeness) can improve the learning experience in interactive scenarios.

Although feedback and demonstration have been largely investigated in reinforcement learning (Silva et al., 2019), the design and evaluation of natural language explanations that foster knowledge transfer in both human-agent and agent-agent scenarios is hardly explored.

Our contribution aims to optimize this knowledge transfer among agents by using explanation-guided exploration. We refer to explanations as the set of information that aims to convey a causality by comparing counterfactuals in the task, i.e, providing the reward that could have been obtained if a different action would have been chosen. Instead of providing the optimal solution for the task, this approach lets the learner infer the best strategy to pursue. In this work, we provide (1) *an overview on the topic of natural language explanations in interactive learning scenarios*, and (2) *a preliminary computational experiment* to evaluate the effect of explanation and demonstration on a learning agent performance in a two-agents setting. We then discuss our results in light of effective natural language explanations for both agents and humans.

2 On Natural Language Explanations in Interactive Learning Scenarios

Humans use the flexibility of natural language to express themselves and provide various forms of feedback, e.g., via counterfactuals. To be successful, artificial agents must therefore be capable of both learning from and using natural language explanations; especially in unstructured environments with human presence. Recent advances in grounded-language feedback state that, although there is a conceptual difference between natural language explanations and tuples that hold information about the environment, natural language is still a favorable candidate for building models that acquire world knowledge (Luketina et al., 2019; Schwartz et al., 2020; Liu and Zhang, 2017; Stiennon et al., 2020). Along this line, training agents to learn

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 61–66,
Dublin, Ireland, 18 December 2020.

rewards from natural language explanations has been widely explored (Sumers et al., 2020; Najar and Chetouani, 2020; Najar et al., 2020; Krening et al., 2017; Knox et al., 2013; Li et al., 2020; Chuang et al., 2020). The interestingness of Sumers et al. (2020) approach lays in grounding the implementation of two artificial agents on a corpus of naturalistic forms of feedback studied in educational research. The authors presented a general method that uses sentiment analysis and contextualization to translate feedback into quantities that reinforcement learning algorithms can reason with. Similarly, (Ehsan and Riedl, 2020) build a training corpus of state-action pairs annotated with natural language explanations with the intent of rationalizing the agent's action or behavior in a way that closely resemble how a human would most likely do.

Existing literature reviews and experimental studies paired natural language feedback with demonstrations of the corresponding tasks to learn the mapping between instructions and actions (Najar and Chetouani, 2020; Taylor, 2018). This aspect has been studied also in the context of real-time interactive learning scenarios in which the guidance and the dialog with a human tutor is often realized by providing explanations (Thomaz et al., 2005; Li et al., 2020).

Following the idea of *AI rationalization* introduced by (Ehsan and Riedl, 2020), our work approaches the generation of explanations as a problem of translation between ad-hoc representations of an agent's behavior and the shape of the reward function. On the contrary, to achieve our goal we use counterfactuals that can be easily encoded in natural language.

2.1 Explanations for Humans

There exists a substantial corpus of research that investigates explanations in philosophy, psychology, and cognitive science. Miller (Miller, 2019) argues that the way humans explain to each other can inform ways to provide explanation in artificial intelligence. In this context, some authors showed that revealing the inner workings of a system can help humans better understand the system. This is often realized by either generating natural language explanations and visualizing otherwise hidden information (Wallkotter, Tulli, Castellano, Paiva, and Chetouani, 2020). Studies on human learning suggest that explanations serve as a guide to generalization. Lombrozo et al. (Lombrozo and Gwynne, 2014) compared the properties of mechanistic and functional explanations for generalizing from known to novel cases. Their results show that the nature of different kinds of explanations can thus provide key insights into the nature of inductive constraints, and the processes by which prior beliefs guide inference.

Above literature highlights the central role of causality in explanation and the vast majority of everyday explanations invoke notions of cause and effect (Keil, 2006). Therefore, we grounded our explanation formalization in this idea of differentiating properties of competing hypothesis (Hoffmann and Magazzeni, 2019) by comparison of contrastive cases (Madumal et al., 2019).

2.2 Explanations for Agents

Several attempts have been made to develop explanations about the decision of an autonomous agent. Many approaches focus on the interpretation of human queries by either mapping inputs to query or instruction templates (Hayes and Shah, 2017; Lindsay, 2019; Krening et al., 2017), by using an encoder-decoder model to construct a general language-based critique policy (Harrison et al., 2018), or by learning structural causal models for identifying the relationships between variables of interest (Madumal et al., 2019).

However, for a model to be considered explainable, it is necessary to account for the observer of the explanation. In this regard, the research of Lage et al. (2019) investigates the effect of the mismatch between the model used to extract a summary of an agent's policy and the model used from another agent to reconstruct the given summary.

Focusing onto experimental work about knowledge transfer between agents, there exist two main approaches to solve this problem: (1) by reusing knowledge from previously solved tasks, (2) by reusing the experience of another agent. The latter is called inter-agent transfer learning, and is often realized thought human feedback, action advising, and learning from demonstration (Argall et al., 2009; Fournier et al., 2019; Jacq et al., 2019). Some authors refer to policy summarization or shaping when the feedback, advice or demonstration summarize the agent's behavior with the objective of transferring information to another agent (Amir and Amir, 2018). Heuristic based approaches extract diverse important states based on state similarity and

q-values, while machine teaching and inverse reinforcement learning approaches extrapolate state-action pairs useful for recovering the agent's reward function (Brown and Niekum, 2018). We take inspiration from policy summarization and learning from demonstration approaches, and extend it by considering explanation-based exploration. Differently from Fournier et al. (2019) and Jacq et al. (2019) we investigate the topic of transfer learning having a two-agents setting and a q-learner. Furthermore, in contrast with the existing approaches that evaluate explanation by measuring the accuracy of an agent's prediction about another agent behavior, we focus on the effect of the explanation on the agent learning.

3 Experiments

To operationalize the constructs discussed above, we have created an interactive learning scenario allowing both human-agent, and agent-agent interaction. We present initial results that use this interactive scenario to compare different kinds of information provided to the learner.

3.1 Hypothesis

We hypothesize that the agent receiving both, explanations and demonstrations, will learn faster than agents that only receive one of these additional forms of teaching signals. Additionally, all three agents will learn faster than an agent learning by itself.

3.2 Materials

Environment The environment is based on Papi's Minicomputer[1], a competitive two-player game, and it enables learning from explanations, demonstrations, and own experience. Papi's Minicomputer is a non-verbal language to introduce children to mechanical and mental arithmetic through decimal notation with binary positional rules. This environment can be taken as an example of a dynamic, navigational environment. Previous studies involving children, used the same environment, and compared optimal and suboptimal actions, giving an information about the effect of those actions in a certain amount of future steps (Tulli et al., 2020).

Learning Agent The learning agent is an agent that chooses moves using a Q-table. It learns from own experience using q-learning ($\alpha = 0.8$, $\gamma = 0.99$) to solve a Markov Decision Process (MDP), in which the optimal Q-value function is $Q*(s,a) = max_\pi Q^\pi(s,a)$ (Sutton and Barto, 2005). Examples from demonstrations are treated in the same way (direct q-learning update). Examples from explanations are converted into a format that allows using a q-learning update by summing the reward from the explainer's actual action with the explained reward difference.

Explainer Agent The explainer agent is model-based and plans moves using the depth limited min-max algorithm with search depth of 3. The agent is also capable of giving demonstrations and explanations (see below).

Demonstrations Demonstrations are additional examples given to the learning agent on top of the self-exploration (plain condition). It allows the agent to learn about states and transitions that it has not explored directly by itself. Concretely, to generate a set of demonstrations, the explainer agent selects 10 random states and generates actions for these states according to its policy. It then uses its task model to compute the corresponding next state and computes the reward obtained by this transition. The explainer then gives this information (state, action, next state, reward) to the learner.

Explanations Similar to demonstrations, explanations are examples given to the learning agent on top of the self-exploration (plain condition). However, differently from demonstrations, explanations contrast alternative actions in the same state and aim to suggest a casual relationship between examples by giving a measure of how good the performed action is.

To generate a set of explanations, the explainer agent first computes the actual action that it will perform in the current state. It computes the next state and the reward associated with this transition. Then, it chooses up to three alternative actions at random and simulates the resulting alternative state and associated reward. Finally, the agent computes the difference between the alternative reward and the reward from the actual action.

All this information (current state, actual action, next state, reward, alternative action, alternative state, reward difference) is then combined and given to the learning agent as an explanation. This is the agent-agent scenario equivalent to a natural language encoding using template sentences. Turned into natural language, such an explanation

[1] http://stern.buffalostate.edu/CSMPProgram/String, consulted on Oct 2020

could take the form of: "I am doing *action* which would give me *reward* and lead to *next state*, because doing *alternative action* would lead to *alternative state* and have *reward difference* points more/less."

3.3 Design

We designed an experiment with four conditions: (1) learning from own experience only [**plain**], (2) learning from experience and demonstrations [**demonstration**], (3) learning from experience and explanations [**explanations**], and (4) learning from experience, demonstrations, and explanations [**both**]. For each condition we let the learning agent play against the explainer agent until it has seen $100,000$ examples in total from any source; i.e., to compute the total number of examples we sum the examples from exploration by itself, from demonstration, and from explanations.

In the condition *plain* the learner agent receives 1 example in each step (self-exploration). In the condition *demonstration*, the learner agent receives 11 examples in each step (1 from self-exploration, 10 from demonstration). In the condition *explanation*, the agent receives up to 4 examples in each step (1 from self-exploration, and up to 3 from explanations, depending on how many alternative actions are available in that state). In the condition *both* the learner agent receives up to 14 examples in each step, one from self-explanation, 10 from demonstrations, and up to 3 from explanations. This means that the number of steps and episodes may differ between conditions, but the total number of samples (i.e., examples) is matched between conditions. This means we are providing the same amount of search-space coverage in each condition.

During a single episode of the game, the learning agent updates its policy at every turn. If it is the learning agent's turn, it performs an update based on its own experience (all conditions). If it is the explainer's turn, the learning agent may receive a set of demonstrations and/or explanations - depending on the condition -, which it uses to update its policy. Then, the learning agent updates its policy again based on the explainer's move (all conditions). The explainer does not update its policy in this setup.

To create a dataset to analyze the performance, we train the agent in each condition for $N = 100$ trials (total of 400 trials). We track the outcome of the game (win/loss) and a rolling average (window size 10) of the current win rate.

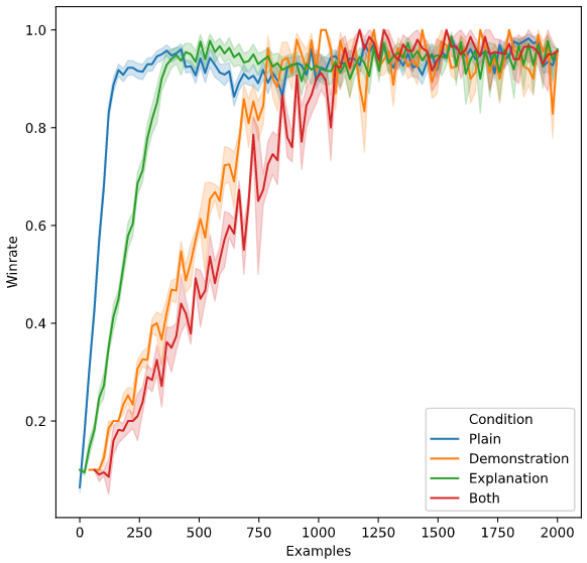

Figure 1: Average (N=100) amount of examples needed to obtain a desired winrate against the explainer agent. The number of examples is calculated as the sum of all examples obtained from self-exploration, demonstrations, and explanations.

3.4 Results

After performing the experiment, we plotted the average number of examples needed for a given winrate grouped by condition (figure 1). The agent begins to perform better than the explainer agent very early in the learning process, which is visualized by a suitable winrate with less than 250 examples. Then, agents from all conditions begin to quickly learn to dominate the explainer agent, with *the agent from the explanation condition requiring the least amount of samples* to win the majority of games. Having access to demonstrations also yields a slight advantage in learning, especially early in the training process. Interestingly, having access to both, demonstrations and explanations, does not lead to improvements.

4 Discussion

In above section we organized the literature on the topic of natural language in interactive learning scenarios involving humans and agents. To date, several excellent works exist on the topic of explainability and natural language technologies, but there it seems to be a gap for experimental work that aims to investigate the concept of explainable AI for transfer learning in both human-agent and agent-agent scenarios.

We expected that the proposed counterfactual structure of an agent's explanations would affect

the learning of another agent interacting in the same environment. Overall, the data did not confirm this hypothesis. We assume that the impact of the formalization of the demonstrations and the explanations is less strong than other learning parameters. Furthermore, the access to both demonstrations and the explanations might have influenced erroneously the agent's reasoning about the task. Future work should consider isolating the problem of comparing different types of information employing other rationale that can be suitable, such as inverse reinforcement learning.

Another challenging future direction is represented by the implementation of methods that model the recipient of an explanation. Inferring the learner understanding of the task through partial observations of its state would help in driving the explainer's selection of informative examples.

One of the aspect we neglected in the current study is more realistic and reactive behaviors on both the part of the learner and the explainer. On this subject, while any given agent may not be an expert during learning, accounting for the explainable agency of agents that are not experts remains a topic of future work.

Using counterfactuals to allow agents to understand the effects of their actions seems a promising approach. However, this is not always applicable in complex environment involving humans. If we consider the Hex Game with a number of states of around 10^{92}, generating counterfactuals in natural language might conduct to probabilistic explanations and increase mental overload, leading to performance degradation.

Considering a training corpus of annotated natural language explanations provided by humans appear to be a necessary requirement to extend our findings to human-agent scenarios. Following the same line, testing the effect of agents' explainability on human learning requires challenging long-term studies. The evaluation framework is, in fact, an open challenge. Further evaluation about the effects of the provided explanations on several metrics beyond the human's performance is needed to support our claims.

5 Conclusion

Throughout this paper, we contextualize natural language explanations with a specific focus on learning scenarios. We gave an overview of the existing literature bridging the concept of explanation in humans and artificial agents and showing that explainability is receiving attention in the context of multi-agent settings. We proposed a preliminary computational experiment for comparing demonstrations and explanations and discuss limitations and future work.

Acknowledgements

We acknowledge the EU Horizon 2020 research and innovation program for grant agreement No 765955 ANIMATAS project. This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020.

References

D. Amir and Ofra Amir. 2018. Highlights: Summarizing agent behavior to people. In *AAMAS*.

Brenna Argall, S. Chernova, M. Veloso, and B. Browning. 2009. A survey of robot learning from demonstration. *Robotics Auton. Syst.*, 57:469–483.

Daniel S. Brown and Scott Niekum. 2018. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *AAAI*.

Y. Chuang, X. Zhang, Yuzhe Ma, M. Ho, Joseph L Austerweil, and Xiaojin Zhu. 2020. Using machine teaching to investigate human assumptions when teaching reinforcement learners. *ArXiv*, abs/2009.02476.

Upol Ehsan and Mark O. Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. *ArXiv*, abs/2002.01092.

P. Fournier, C. Colas, M. Chetouani, and O. Sigaud. 2019. Clic: Curriculum learning and imitation for object control in non-rewarding environments. *IEEE Transactions on Cognitive and Developmental Systems*, pages 1–1.

Brent Harrison, Upol Ehsan, and Mark O. Riedl. 2018. Guiding reinforcement learning exploration using natural language. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 1956–1958, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Bradley Hayes and Julie A. Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 303–312, New York, NY, USA. Association for Computing Machinery.

Jörg Hoffmann and Daniele Magazzeni. 2019. Explainable ai planning (xaip): Overview and the case of contrastive explanation (extended abstract). In *Reasoning Web*.

Alexis Jacq, Matthieu Geist, Ana Paiva, and Olivier Pietquin. 2019. Learning from a learner. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2990–2999, Long Beach, California, USA. PMLR.

F. Keil. 2006. Explanation and understanding. *Annual review of psychology*, 57:227–54.

W. B. Knox, P. Stone, and C. Breazeal. 2013. Training a robot via human feedback: A case study. In *ICSR*.

S. Krening, B. Harrison, K. M. Feigh, C. L. Isbell, M. Riedl, and A. Thomaz. 2017. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55.

Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. 2019. Exploring computational user models for agent policy summarization. *CoRR*, abs/1905.13271.

Toby Jia-Jun Li, Tom Mitchell, and Brad Myers. 2020. Interactive task learning from GUI-grounded natural language instructions and demonstrations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 215–223, Online. Association for Computational Linguistics.

Alan Lindsay. 2019. Towards exploiting generic problem structures in explanations for automated planning. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, page 235–238, New York, NY, USA. Association for Computing Machinery.

Rui Liu and X. Zhang. 2017. A review of methodologies for natural-language-facilitated human–robot cooperation. *International Journal of Advanced Robotic Systems*, 16.

T. Lombrozo and Nicholas Z. Gwynne. 2014. Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, 8.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, E. Grefenstette, S. Whiteson, and Tim Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. *ArXiv*, abs/1906.03926.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. Explainable reinforcement learning through a causal lens. *ArXiv*, abs/1905.10958.

T. Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *ArXiv*, abs/1706.07269.

A. Najar, Olivier Sigaud, and M. Chetouani. 2020. Interactively shaping robot behaviour with unlabeled human instructions. *Autonomous Agents and Multi-Agent Systems*, 34:1–35.

Anis Najar and Mohamed Chetouani. 2020. Reinforcement learning with human advice. a survey.

E. Schwartz, Guy Tennenholtz, Chen Tessler, and S. Mannor. 2020. Language is power: Representing states using natural language in reinforcement learning. *arXiv: Computation and Language*.

Felipe Leno Da Silva, Garrett Warnell, Anna Helena Reali Costa, and Peter Stone. 2019. Agents teaching agents: a survey on inter-agent transfer learning. *Autonomous Agents and Multi-Agent Systems*, 34:1–17.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback.

Theodore R. Sumers, M. Ho, R. D. Hawkins, K. Narasimhan, and T. Griffiths. 2020. Learning rewards from linguistic feedback. *ArXiv*, abs/2009.14715.

R. Sutton and A. Barto. 2005. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 16:285–286.

M. Taylor. 2018. Improving reinforcement learning with human input. In *IJCAI*.

A. Thomaz, Guy Hoffman, and C. Breazeal. 2005. Real-time interactive reinforcement learning for robots. In *American Association for Artificial Intelligence*.

Silvia Tulli, Marta Couto, Miguel Vasco, Elmira Yadollahi, Francisco Melo, and Ana Paiva. 2020. Explainable agency by revealing suboptimality in child-robot learning scenarios. In *Social Robotics*, pages 23–35, Cham. Springer International Publishing.

Sebastian Wallkotter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2020. Explainable agents through social cues: A review.

Generating Explanations of Action Failures in a Cognitive Robotic Architecture

Ravenna Thielstrom, Antonio Roque, Meia Chita-Tegmark, Matthias Scheutz
Human-Robot Interaction Laboratory
Tufts University
Medford, MA 02155
{ravenna.thielstrom,antonio.roque,mihaela.chita_tegmark,matthias.scheutz}@tufts.edu

Abstract

We describe an approach to generating explanations about why robot actions fail, focusing on the considerations of robots that are run by cognitive robotic architectures. We define a set of Failure Types and Explanation Templates, motivating them by the needs and constraints of cognitive architectures that use action scripts and interpretable belief states, and describe content realization and surface realization in this context. We then describe an evaluation that can be extended to further study the effects of varying the explanation templates.

1 Introduction

Robots that can *explain* why their behavior deviates from user expectations will likely benefit by better retaining human trust (Correia et al., 2018; Wang et al., 2016). Robots that are driven by a cognitive architecture such as SOAR (Laird, 2012), ACT-R (Ritter et al., 2019), or DIARC (Scheutz et al., 2019) have additional requirements in terms of connecting to the architecture's representations such as its belief structures and action scripts. If properly designed, these robots can build on the interpretability of such architectures to produce explanations of action failures.

There are various types of cognitive architectures, which may be defined as "abstract models of cognition in natural and artificial agents and the software instantiations of such models" (Lieto et al., 2018) but in this effort we focus on the type that uses action scripts, belief states, and natural language to interact with humans as embodied robots in a situated environment. In Section 2 we describe an approach to explaining **action failures**, in which a person gives a command to a robot but the robot is unable to complete the action. This approach was implemented in a physical robot with a cognitive architecture, and tested with a preliminary evaluation as described in Section 3. After comparing our effort to related work in Section 4, we finish by discussing future work.

2 An Approach to Action Failure Explanation

Our approach is made up of a set of Failure Types, a set of Explanation Templates, algorithms for Content Realization, and algorithms for Surface Realization.

2.1 Failure Types

We have defined an initial set of four different failure types, which are defined by features that are relevant to cognitive robots in a situated environment. One approach to designing such robots is to provide a database of action scripts that it knows how to perform, or that it is being taught how to perform. These scripts often have prerequisites that must be met before the action can be performed; for example, that required objects must be available and ready for use. These action scripts also often have defined error types that may occur while the action is being executed, due to the unpredictability of the real world. Finally, in open-world environments robots usually have knowledge about whether a given person is authorized to command a particular action. Incorporating these feature checks into the architecture of the robot allows for automatic error type retrieval when any of the checks fail, essentially providing a safety net of built-in error explanation whenever something goes wrong. These features are used to define the failure types as follows. When a robot is given a command, a series of checks are performed.

First, for every action necessary to carry out that command, the robot checks to see whether the action exists as an action script in the robot's database of known actions. If it does not, then the action is not performed due to an **Action Ignorance** failure

Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020), pages 67–72,
Dublin, Ireland, 18 December 2020.

type. This would occur in any situation where the robot lacks knowledge of *how* to perform an action, for example, if a robot is told to walk in a circle, but has not been instructed what walking in a circle means in terms of actions required.

Second, the robot checks whether it is obligated to perform the action, given its beliefs about the authorization level of the person giving the command. If the robot is not obligated to perform the action, the system aborts the action with an **Obligation Failure** type. An example of this failure would be if the person speaking to the robot does not have security clearance to send the robot into certain areas.

Third, the robot checks the conditions listed at the start of the action script, which define the facts of the environment which must be true before the robot can proceed. The robot evaluates their truth values, and if any are false, the system exits the action with a **Condition Failure** type. For example, a robot should check prior to walking forward that there are no obstacles in its way before attempting that action.

Otherwise, the robot proceeds with the rest of the action script. However, if at any point the robot suffers an internal error which prevents further progress through the action script, the system exits the action with a **Execution Failure** type. These failures, in contrast, to the pre-action condition failures, come during the execution of a primitive action. For example, if a robot has determined that it is safe to walk forward, but after engaging its motors to do just that, either an internal fault with the motors or some other unforseen environmental hazard result in the motors not successfully engaging. In either case, from the robot's perspective, the only information it has is that despite executing a specific primitive (engaging the motors), it did not successfully return the expected result (motors being engaged).

2.2 Explanation Templates

Once the type of failure is identified, the explanation assembly begins. The basic structure of the explanation is guided by the nature of action scripts. We consider an inherently interpretable action representation that has an intended goal **G** and failure reason **R** for action **A**, and use these to build four different explanation templates of varying depth.

The **GA** template captures the simplest type of explanation: "I cannot achieve *G* because I cannot do *A*." For example, "I cannot *prepare the product* because I cannot *weigh the product*."

The **GR** template captures a variant of the first explanation making expicit reference to a reason: "I cannot achieve *G* because of *R*." For example, "I cannot *prepare the product* because *the scale is occupied*."

The **GGAR** template combines the above two schemes by explicitly linking *G* with *A* and *R*: "I cannot achieve *G* because to achieve *G* I must do *A*, but *R* is the case." For example, "I cannot *prepare the product* because to *prepare something* I must *weigh it*, but *the scale is occupied*."

Finally, the **GGAAR** template explicitly states the goal-action and action-failure reason connections: "I cannot achieve *G* because for me to achieve *G* I must do *A*, and I cannot do *A* because of *R*." For example, "I cannot *prepare the product* because to *prepare something* I must *weigh it*, and I cannot *weigh the product* because *the scale is occupied*."

2.3 Content Realization

Given the failure type that has occurred, and the explanation template (which is either set as a parameter at launch-time or determined at run-time), a data structure carrying relevant grammatical and semantic information is constructed.

The code version of an explanation template contains both bound and generic variables, which in the GGAAR template looks like:

```
can(not(BOUND-G),
    because(advrb(infinitive(GENERIC-G),
    must(GENERIC-A)), can(not(BOUND-A),
    because(REASON))))
```

`BOUND-G` and `GENERIC-G` are the bound and unbound versions of the goal. For example `did(self,prepare(theProduct))` is the bound version which specifies the product, and `did(self,prepare(X))` is the unbound version.

Similarly, `GENERIC-A` is the generic form of the sub-action which failed, such as `did(self,weigh(X))`, `BOUND-A` is the lowest-level sub-action, such as `did(self, weigh(theProduct))`, and `REASON` is the error reason, such as `is(theScale,occupied)`.

So the resulting form would look like:

```
can(not(prepare(self,theProduct)),
    because(advrb(infinitive(
    prepare(self,X)), must(
    weigh(self,X))),
    can(not(weigh(self,theProduct)),
    because(is(theScale,occupied)))))
```

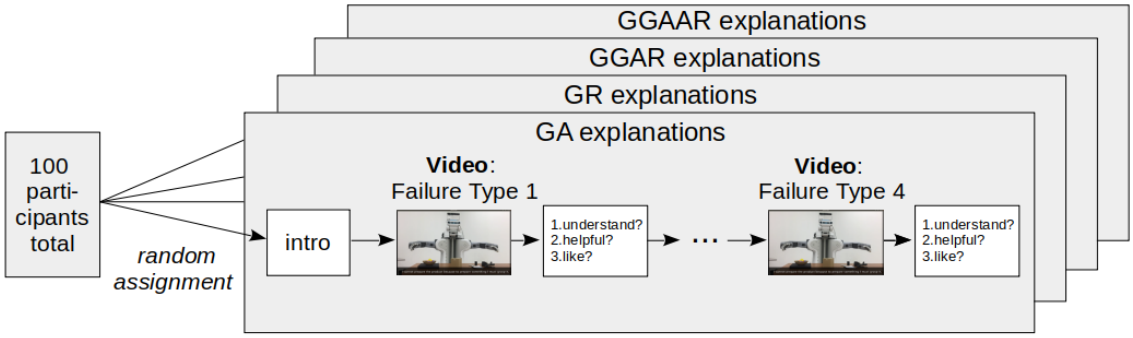

Figure 1: Study Procedure.

and would then be submitted to the Surface Realization process.

2.4 Surface Realization

Translating the semantic form of the explanation into natural language is a matter of identifying grammatical structures such as premodifiers, infinitives, conjunctions, and other parts of speech by recursively iterating through the predicate in search of grammar signifiers.

This process involves populating grammatical data structures (i.e. clauses) with portions of the semantic expression and their relevant grammatical information. During each recursive call, the name of the current term is checked to see if it matches a grammatical signifier; if so, it is unwrapped further and recurses over the inner arguments. Without any more specific signifiers, the term name can be assumed to be a verb, the first argument the subject, and the second the object of the clause. The grammatical signifiers are used to assign grammatical structure as needed, which are then conjugated and fully realized using SimpleNLG (Gatt and Reiter, 2009) into natural language, such as: "I cannot prepare the product because to prepare something I must weigh it, and I cannot weigh the product because the scale is occupied."

3 Evaluation

To validate our system, we conducted a user study. Besides testing the components all working together, we were also interested in understanding the effect of the different types of explanation templates on human perceptions of the explanations given. This study was conducted under the oversight of an Institutional Review Board.

3.1 Methods

100 participants were recruited via Amazon's Mechanical Turk and completed this study online through a web interface.

As shown in Figure 1, after a brief introduction, participants were shown four different videos, one at a time, in which a robot was instructed to "prepare the product." In each video the robot explained that it could not complete the task due to one of four failure types described in Section 2.1. For example, in the first video the robot might explain that it did not know how to perform the action, in the second video the robot might explain that the person was not authorized to make the action request, in the third video the robot might explain that the scale was occupied, and in the fourth video the robot might explain that their pathfinding algorithm had failed. 25 participants were shown videos in which the explanations used the GA template, 25 in which the videos used the GR template, 25 with the GGAR template, and 25 with the GGAAR template.

After each video the participants were asked three questions.

First, to assess their understanding of how the robot failed its task, the participants were asked "What would you do in order to allow the robot to complete the task?" and were given 5 possible solutions in a multiple-choice format, only one of which was correct. For example, given the Condition Failure error explanation in the GGAAR format: "I cannot prepare the product because to prepare the product I must weigh it, and I cannot weigh the product because the scale is occupied" possible solutions are: (1) I would have the robot learn how to weigh things, (2) I would have the robot's pathfinding component debugged, (3) I would clear the scale, (4) I would move the scale closer to the

Figure 2: Screen Capture from Example Video with Generated Text. A robot, given an instruction, explains an action failure.

robot, (5) I would have the robot's vision sensors repaired, where 3 is the correct solution.

Second, the participants were asked "How helpful was the robot's explanation?" on a 5-point Likert scale where 1 was "Not at all" and 5 was "Extremely."

Third, the participants were asked "How much did you like the robot's explanation?" on a 5-point Likert scale where 1 was "Not at all" and 5 was "Extremely."

These questionnaire items were selected with a focus on the social interaction between the robot and the human rather than the fluency or semantic meaning of the natural language generation itself. Perceived helpfulness and likability are both metrics of trust in a human-robot interaction, and more specifically, they are indications of the human being comfortable cooperating with the robot. Thus we aimed to assess how well the robot's explanation communicated the problem to the human (with the accuracy questions), in addition to how successful the explanations were as a social interaction.

The failure explanations in the videos were generated using a Wizard-of-Oz approach. Our explanation approach was implemented in a PR2 robot using the DIARC cognitive architecture (Scheutz et al., 2019). We filmed a PR2 robot performing preparatory-type movement (looking down at a table full of miscellaneous items, raising its hands, looking back up at the camera) before halting and delivering an audio failure explanation report (generated by our system as described in Section 2 and recorded separately, then edited into the video along with subtitles.) A screen capture of an example video is shown in Figure 2. An example video of an explanation is located here:

https://youtu.be/2j7r1S6zT90

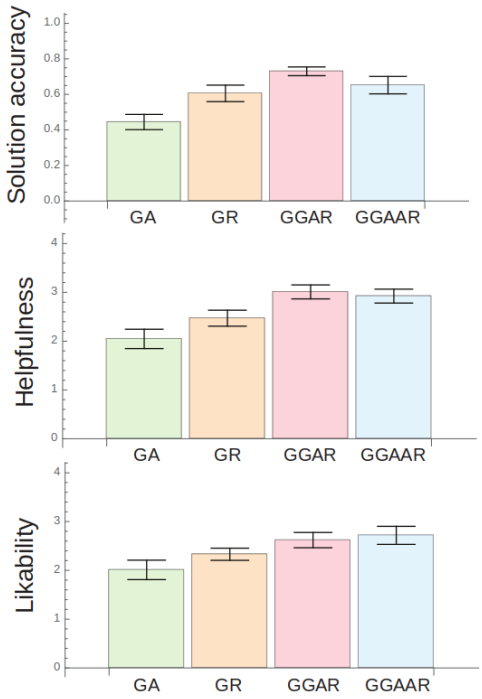

Figure 3: Evaluation Results. Proportion of accurate responses, and Likert-scale ratings of likability and helpfulness, based on Explanation Template.

3.2 Results

To investigate how the different explanation schemas the robot gave allowed the participants to select the **accurate solution** for fixing the problem, we conducted a one-way ANOVA with the solution accuracy (number of correct solutions selected across 4 different error types) as our dependent variable, and explanation template (GA, GR, GGAR and GGAAR) as the independent variable. We observed a significant effect of explanation template on solution accuracy $F(3, 97) = 8.61$, $p < .001$, η_p^2=.21. Further pairwise comparisons with Tukey-Kramer corrections revealed that GA explanations lead to significantly lower solution accuracy than GGAAR ($p = .004$), GAR ($p < .001$) and GR ($p = .031$) explanations. No other significant differences between explanation templates were observed. In other words, short explanations lacking a reason for failure will result in decreased understanding of how to best address the failure.

We then studied perceived **explanation helpfulness**. We conducted a one-way ANOVA with explanation helpfulness as the dependent variable and explanation template (GA, GR, GGAR, GGAAR) as the independent variable. We found a significant effect of explanation template, $F(3, 97) = 7.34$, $p < .001$, η_p^2=.30. Pairwise comparisons revealed

a similar pattern of results as for solution accuracy: participants perceived the GA explanations to be less helpful than GGAAR ($p = .002$) and GGAR ($p < .001$), however, unlike the solution accuracy no significant differences were found between GA-type explanations and GR-type ones. No other significant differences in helpfulness were found between explanation.

Finally, we investigated **explanation likability** by conducting a one-way ANOVA with explanation likability as the dependent variable and explanation template (GA, GR, GGAR, GGAAR) as the independent variable. We found again a significant main effect of explanation schema $F(3, 96) = 3.59$, $p = .016$, η_p^2=.10. Pairwise comparisons revealed that GA explanations were liked less than GGAAR ($p = 0.021$) and GGAR ($p = 0.053$) but not significantly different from GR. We found no other significant differences in perceived likability between explanation templates.

This study highlights the value of providing a failure reason R in the explanation templates, which is shown by the reduced measures of the GA explanations.

4 Related Work

Human-Robot Interaction (HRI) research on explaining the actions of robots (Anjomshoae et al., 2019) is related to research on explaining planning decisions (Fox et al., 2017; Krarup et al., 2019), on generating language that describes the pre- and post-conditions of actions in planners (Kutlak and van Deemter, 2015), and on generating natural language explanations from various types of meaning representations (Horacek, 2007; Pourdamghani et al., 2016).

In HRI work that focuses on error reporting, Briggs and Scheutz (2015) defined a set of *felicity conditions* that must hold for a robot to accept a command. They outlined an architecture that reasons about whether each felicity condition holds, and they provided example interactions, although they did not evaluate an implementation of their approach. Similarly, Raman et al. (2013) used a logic-based approach to identify whether a command can be done, and provided example situations, but no evaluation. Our approach is similar in that we define a set of failure types for action commands, but we implement and evaluate our approach with a user study. Other recent HRI work has included communicating errors using non-verbal actions to have a robot express its inability to perform an action (Kwon et al., 2018; Romat et al., 2016), which does not focus on more complex system problems using natural language communications as we do.

There has also been recent work on user modeling and tailoring responses to users in robots (Torrey et al., 2006; Kaptein et al., 2017; Sreedharan et al., 2018). In one effort worth building upon, Chiyah Garcia et al. (2018) used a human expert to develop explanations for unmanned vehicle decisions. These explanations followed Kulesza et al. (2013) in being characterized in terms of *soundness*, relating the depth of details, and *completeness*, relating to the number of details. Chiyah Garcia et al. found links between the "low soundness and high completeness" condition and intelligibility and value of explanations.

5 Conclusions

We have described an approach to generating action failure explanations in robots, focusing on the needs and strengths of a subset of cognitive robotic architectures. This approach takes advantage of the interpretability of action scripts and belief representations, and is guided by recent directions in HRI research. Importantly, the explanation of this approach is not a post-hoc interpretation of a black-box system, but is an accurate representation of the robot's operation.

Various aspects of the approach are being continually refined. Currently, new Failure Types are being investigated, and the content realization and surface realization algorithms are being revised and tested.

Finally, the evaluation in Section 2.2 describes a preliminary approach to comparing the relative impact of the various explanation templates. We are pursuing additional studies focusing on varying the explanations produced. Initial studies would be video-based, after which follow-up studies would be conducted in the context of a task being performed either in person, or via a virtual interface that we have constructed, and the goal would be to examine the ways that context features such as user model, physical setting, and task state affect the type of explanation required.

Acknowledgment

We are grateful to the anonymous reviewers for their helpful comments. This work was in part funded by ONR grant #N00014-18-2503.

References

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*.

Gordon Michael Briggs and Matthias Scheutz. 2015. "Sorry, I Can't Do That": Developing mechanisms to appropriately reject directives in human-robot interactions. In *2015 AAAI fall symposium series*.

Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation*.

Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S. Melo, and Ana Paiva. 2018. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18.

Maria Fox, Derek Long, and Daniele Magazzeni. 2017. Explainable planning. In *Proceedings of the IJCAI 2017 Workshop on Explainable AI*.

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.

Helmut Horacek. 2007. How to build explanations of automated proofs: A methodology and requirements on domain representations. In *Proceedings of AAAI ExaCt: Workshop on Explanation-aware Computing*, pages 34–41.

Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 676–682. IEEE.

Benjamin Krarup, Michael Cashmore, Daniele Magazzeni, and Tim Miller. 2019. Model-based contrastive explanations for explainable planning. In *Proceedings of the ICAPS 2019 Workshop on Explainable Planning (XAIP)*.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10. IEEE.

Roman Kutlak and Kees van Deemter. 2015. Generating Succinct English Text from FOL Formulae. In *Procs. of First Scottish Workshop on Data-to-Text Generation*.

Minae Kwon, Sandy H Huang, and Anca D Dragan. 2018. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 87–95. ACM.

John E Laird. 2012. *The Soar cognitive architecture*. MIT press.

Antonio Lieto, Mehul Bhatt, Alessandro Oltramari, and David Vernon. 2018. The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research*, 48:1 – 3.

Nima Pourdamghani, Kevin Knight, and Ulf Hermjakob. 2016. Generating English from abstract meaning representations. In *Proceedings of the 9th International Natural Language Generation conference*, pages 21–25.

Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton CT Lee, Mitchell P Marcus, and Hadas Kress-Gazit. 2013. Sorry Dave, I'm afraid I can't do that: Explaining unachievable robot tasks using natural language. In *Robotics: Science and Systems*, volume 2, pages 2–1.

Frank E Ritter, Farnaz Tehranchi, and Jacob D Oury. 2019. ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3):e1488.

Hugo Romat, Mary-Anne Williams, Xun Wang, Benjamin Johnston, and Henry Bard. 2016. Natural human-robot interaction using social cues. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*.

Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. 2018. Hierarchical expertise level modeling for user specific contrastive explanations. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

Cristen Torrey, Aaron Powers, Matthew Marge, Susan R Fussell, and Sara Kiesler. 2006. Effects of adaptive robot dialogue on information exchange and social relations. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*.

Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *11th ACM/IEEE International Conference on Human-Robot Interaction*.