1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020)

Proceedings

29-30 August, 2020
Santiago de Compostela, Spain

Preface

Welcome to the Proceedings of the 1^{st} Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020)!

This Doctoral Consortium (DC) is under the umbrella of the Mentoring and Communication for Starting Researchers (MC4SR) ECAI Program. In addition to the DC, MC4SR included other events such as Meeting with a EurAI Fellow, Job Fair, and the 9^{th} European Starting AI Researchers' Symposium (STAIRS).

The DC-ECAI 2020 provides a unique opportunity for PhD students, who are close to finishing their doctorate research, to interact with experienced researchers in the field. Senior members of the community are assigned as mentors for each group of students based on the student's research or similarity of research interests. The DC-ECAI 2020, which is held virtually this year, allows students from all over the world to present their research and discuss their ongoing research and career plans with their mentor, to do networking with other participants, and to receive training and mentoring about career planning and career options.

The Programme includes two main sessions:

- Training Session (August 29^{th}). The focus of this session is on training transversal and communication skills. The session has 60 attendants who will learn how to present their work in oral presentations with an emphasis in the practical case of a scientific elevator pitch for both the academy and industry. The session is led by Martin Holm Mortensen (FundingBox Research).
- Mentoring Session (August 30^{th}). This session is aimed at dissemination of the ongoing research developed by Starting AI Researchers. We have 40 speakers who will present and leverage their work to the AI community with brief 3-minute video presentations. Peers and senior AI researchers will attend discussion sessions and provide speakers with feedback and guidance on future research directions. The three best presentations are to be awarded.

We want to thank all attendants and especially to speakers and mentors for contributing to this Doctoral Consortium.

Ulises Cortés and José M. Alonso
DC-ECAI 2020 Chairs

Doctoral Consortium Chairs:

Prof. Ulises Cortés, Barcelona Supercomputing Center and Univ. Politècnica de Catalunya
Dr. José M. Alonso, CiTIUS, University of Santiago de Compostela

Publication Chair:

Dr. Alejandro Catala, CiTIUS, University of Santiago de Compostela

Awards Committee:

Prof. Ramón López de Mántaras, IIIA-CSIC
Prof. Virginia Dignum, Umea University
Prof. Luis Magdalena, Unversidad Politécnica de Madrid
Prof. Ulises Cortés, Barcelona Supercomputing Center and Univ. Politècnica de Catalunya
Dr. José M. Alonso, CiTIUS, University of Santiago de Compostela

Mentors:

Prof. Karina Gibert, Univ. Politècnica de Catalunya
Prof. Ulises Cortés, Barcelona Supercomputing Center and Univ. Politècnica de Catalunya
Dr. Pedro Meseguer, IIIA-CSIC
Dr. Jaume Manero, Barcelona Supercomputing Center
Dr. Manuel Mucientes Molina, CiTIUS, University of Santiago de Compostela
Dr. Manuel Fernández Delgado, CiTIUS, University of Santiago de Compostela
Dr. Alejandro Catala, CiTIUS, University of Santiago de Compostela
Dr. José M. Alonso, CiTIUS, University of Santiago de Compostela

Lecturer:

Martin Holm Mortensen, FundingBox Research

Doctoral Consortium - Training

The Training part in the DC-ECAI 2020 takes the form of a webinar that is split into two sessions to be held by August 29^{th}. These sessions are devoted to training transversal and communication skills. The attendants are expected to learn how to present their work in oral presentations with an emphasis in the practical case of a scientific elevator pitch for both the academy and industry.

The session is led by Martin Holm Mortensen who is Senior Project Manager at FundingBox Research; Master of Science in European & International Politics from The University of Edinburgh and Ba. in Public Administration from Roskilde University. He has 20 years of experience in startup management as a founder, c-level manager and on the Board of startups like ComputerPeople (Nordics), CapaSystems (USA + Denmark), CopenhagenDowntown (Denmark) and Clippingpath.com (UK + Bangladesh). During the last 7 years Martin has changed focus towards helping startups with commercialization, sales and investments externally as a Project Manager, Consultant and Investment Manager in some of Europa's best accelerators and growth projects like the Fundingbox lead EU program "IMPACT Growth", the Nordic accelerator "Accelerace Management" and the UK accelerator "NBG" in Nottingham.

Doctoral Consortium - Mentoring

The Mentoring part in the DC-ECAI 2020 is organized into five Mentoring Tracks (MT), to be held by August 30^{th}. MTs have chairs who lead the session, introducing each speaker and handling questions and answering. There are 8 speakers in each track. Each speaker has a slot of 15 min for a brief personal introduction which is followed by an informal conversation regarding current and future research work.

- MT1: Knowledge Representation and Reasoning (Chair: Karina Gibert)
  - Munyque Mittelmann, "Auction Description Language (ADL): a general framework for representing auction-based markets"
  - Periklis Mantenoglou, "Online probabilistic interval-based event calculus"
  - Fernando E. Casado, "Learning from the individuals and the crowd in a society of devices"
  - Gianluca Zaza, "Monitoring cardiovascular risk by facial video processing and fuzzy rules"
  - Kenneth Skiba, "Towards a ranking-based semantics using system Z"
  - Ignacio Huitzil, "Advanced management of fuzzy semantic information"
  - Gönül Ayci, "Handling uncertainty and risk in privacy prediction"
  - Marcelo de Souza, "On the automatic configuration of algorithms"
- MT2: Agents, Planning and Scheduling (Chair: Jaume Manero)
  - Francisco J. Gil-Gala, "Evolving priority rules for scheduling problems by means of hyper-heuristics"
  - Justin Svegliato, "A metareasoning framework for planning and execution in autonomous systems"
  - Marcel Tiator, "Smart object segmentation to enhance the creation of interactive environments"
  - Norah Aldahash, "Improving agent interaction safety through environmental control and regulation"
  - Stanislav Sitanskiy, "Recognition and learning of agent behaviour models"
  - Gabriele Sartor, "Integrating open-ended learning and planning for long-term autonomy"
  - Jeferson José Baqueta, "A computational trust model for task delegation"
  - Anas Shrinah, "Verification and validation of planning-based autonomous systems"
- MT3: Natural Language Technology (Chairs: Jose M. Alonso and Alejandro Catala)
  - Kervadec Corentin, "Improving visual reasoning in visual question answering systems"
  - Weilai Xu, "Stylistic dialogue generation in narratives"
  - Aaron Keesing, "Improving robustness of emotion recognition in multi-accent and multilingual speech"
  - Andrea Cascallar Fuentes, "Fuzzy quantified protoforms for data-to-text systems: a new model with applications"
  - Chaina Santos Oliveira, "A two-level item response theory model to evaluate automatic speech recognition systems"
  - Damián Ariel Furman, "Hate speech analysis via argumentation schemes mining"
  - Yago Fontenla-Seco, "A framework for the automatic generation of natural language descriptions of processes"

  - Carlos Andres Lopez Jaramillo, "A model for adopting the omnichannel strategy from a context-aware computing and natural language processing approach"

- MT4: Explainable and Trustworthy AI (Chairs: Pedro Meseguer and Ulises Cortés)
  - Gunay Kazimzade, "Structures and infrastructures around AI: unbiasedness, politics and metrics in data-driven socio-technical systems"
  - Ilse Verdiesen, "Accountability and control over autonomous weapon systems: a framework for comprehensive human oversight"
  - Joyjit Chatterjee, "Explainable AI for intelligent decision support in operations and maintenance of wind turbines"
  - Raúl Antonio del Águila Escobar, "OBOE: an explainable text classification framework"
  - Ettore Mariotti, "Understandable deep learning analysis for very high energy astroparticle physics"
  - Conor Hennessy, "Explaining Bayesian networks in natural language"
  - Ilia Stepin, "Argumentation-based interactive factual and counterfactual explanation generation"
  - Lucas Morillo-Méndez, "Age-related individual differences and their implications for the design of cognitive interactive systems"

- MT5: Big Data and Deep Machine Learning (Chairs: Manuel Mucientes Molina and Manuel Fernández Delgado)
  - Aiste Gerybaite, "Big data in IoE: investigating IT approaches to big data in healthcare whilst ensuring the competing interests of the right to health and the right to privacy"
  - Jorge García-González, "Deep learning neural networks to detect anomalies in video sequences"
  - Russa Biswas, "Embedding based link prediction for knowledge graph completion"
  - Ana Vieira, "Using machine learning for the personalized healthcare of vascular diseases"
  - Hemant Rathore, "Adversarial attacks on Android malware detection models using reinforcement learning"
  - Noelia Rico, "Ranking rules as a tool for reducing the impact of the distance in machine learning methods"
  - Armand Vilalta, "Semantic embeddings in deep convolutional neural networks"
  - Bartosz Piotrowski, "Reinforcement learning for saturation-based theorem provers"

The 40 speakers had provided previously brief 3-minute video presentations which were evaluated by the Awards Committee in terms of:

- Clarity: Did the presentation clearly describe the impact and/or results of the research, including conclusions and outcomes?
- Significance: Is the Thesis significant to the field? Are the communicated results coherent?
- Communication: Did the presenter capture and maintain their audience's attention?

Two weeks in advance to the DC-ECAI 2020 Mentoring session, the Awards Committee screened the 10 finalists in the DC-ECAI 2020 Best Presentation Contest:

- Periklis Mantenoglou (NCSR Demokritos)
- Corentin Kervadec (Universite de Lyon)
- Fernando E. Casado (CiTIUS-USC)
- Ettore Mariotti (CiTIUS-USC)
- Norah Aldahash (University of York)
- Weilai Xu (Bournemouth University)
- Gabriele Sartor (Universita degli Studi di Torino)
- Ignacio Huitzil (University of Zaragoza)
- Noelia Rico (University of Oviedo)
- Armand Vilalta (Barcelona Supercomputing Center)

The three winners will be announced in the DC-ECAI 2020 closing session. In addition to a diploma, the winners will be awarded with a 3-months research stay at CiTIUS-USC, including travel and accommodation expenses, as well as 700 eur per month for subsistence allowance.

Doctoral Consortium - Program (*)

(*) Time is expressed in GMT+2/CEST.

August 29th, 2020

9:00–9:10 **Welcome and Opening**

9:10–10:30 **DC Training (Session 1)**

10:30–10:45 **Break**

10:45–12:15 **DC Training (Session 2)**

August 30th, 2020

13:45–14:00 **Introduction to Mentoring**

14:00–15:00 **DC Mentoring (Session 1)**

14:00–14:15 Munyque Mittelmann (MT1)

Francisco J. Gil-Gala (MT2)

Kervadec Corentin (MT3)

Gunay Kazimzade (MT4)

Aiste Gerybaite (MT5)

14:15–14:30 Periklis Mantenoglou (MT1)

Justin Svegliato (MT2)

Weilai Xu (MT3)

Ilse Verdiesen (MT4)

Jorge García-González (MT5)

14:30–14:45 Fernando E. Casado (MT1)

Marcel Tiator (MT2)

Aaron Keesing (MT3)

Joyjit Chatterjee (MT4)

Russa Biswas (MT5)

14:45–15:00 Gianluca Zaza (MT1)

Norah Aldahash (MT2)

Andrea Cascallar Fuentes (MT3)

Raúl Antonio del Águila Escobar (MT4)

Ana Vieira (MT5)

15:00–15:30 Break

15:30–16:30 DC Mentoring (Session 2)

15:30–15:45 Kenneth Skiba (MT1)

Stanislav Sitanskiy (MT2)

Chaina Santos Oliveira (MT3)

Ettore Mariotti (MT4)

Hemant Rathore (MT5)

15:45–16:00 Ignacio Huitzil (MT1)

Gabriele Sartor (MT2)

Damián Ariel Furman (MT3)

Conor Hennessy (MT4)

Noelia Rico (MT5)

16:00–16:15 Gönül Ayci (MT1)

Jeferson José Baqueta (MT2)

Yago Fontenla-Seco (MT3)

Ilia Stepin (MT4)

Armand Vilalta (MT5)

16:15–16:30 Marcelo de Souza (MT1)

Anas Shrinah (MT2)

Carlos Andrés López Jaramillo (MT3)

Lucas Morillo-Méndez (MT4)

Bartosz Piotrowski (MT5)

16:30–17:00 Concluding Remarks and Farewell

Table of Contents

Auction Description Language (ADL): a General Framework for Representing Auction-based Markets

Munyque Mittelmann[1]

Abstract. The goal of this study is to present a language for representing and reasoning about the rules governing an auction-based market. Such language is at first interest as long as we want to build up digital market places based on auction, a widely used framework for automated transactions. Auctions may differ in several aspects and this variety prevents an agent to easily switch between different (auction-based) markets. The first requirement for building such agents is to have a general language for describing auction-based markets. Second, this language should also allow the reasoning about the key issues of a specific market, namely the allocation and payment rules. To do so, we define a language in the spirit of the *Game Description Language* (GDL): the *Auction Description Language* (ADL) is the first language for describing auctions in a logical framework. ADL is appropriate to represent different types of well-known auctions, such as the English Auction and the Multi-Unit Vickrey Auction. ADL allows us to derive properties about auction protocols (e.g. the protocol finiteness) and the behavior of a rational bidder.

1 INTRODUCTION

A huge volume of goods and services are sold through auctions. Typically, an auction-based market is described by a set of rules stating what are the available actions to the participants, how the winner is determined, and what price should be paid by the winner. There are variants where multiple winners could be considered or payment may also concern the losers. Actually, an Auction protocol may differ in numerous aspects: single or double-side, ascending or descending, single or multi-unit goods, and so on [5].

This great variety of auction protocols prevents any autonomous agent to easily switch between different auction based-markets [10]. Having a language for describing auctions from a general perspective is then at first interest. This language should also allow the reasoning about the key issues of a specific market, namely the allocation and payment rules. Participants may be able to process the auction definition and, consequently, define their bids wrt. these rules.

The goal of this study is to present a logical language for representing and reasoning about the rules governing an auction-based market. Our approach is based on the Game Description Language (GDL) which is a logic-based language for representing and reasoning about game rules; GDL is the official language for the *General Game Playing* challenge [3]. We revisit the GDL variant proposed in [9] and define the logical *Auction Description Language*: we allow numerical variables, comparison and parameters at the opposite of GDL. Handling numerical values is critical for defining the payment and allocation rules.

[1] Université de Toulouse - IRIT, France, e-mail: munyque.mittelmann@irit.fr

1.1 Related work

An auction protocol is an allocation procedure. Its main characteristics are (i) it is a centralized procedure, i.e. it has a central authority (the auctioneer), and (ii) it has monetary transfer. This is not always true for an allocation procedure. For instance, a negotiation protocol may be defined in a distributed (decentralized) approach, where the allocation is the result of a sequence of local negotiation steps [2]. By the other hand, a protocol for exchanging goods or services is not dependant on monetary transfer.

To the best of our knowledge, almost all contributions on the computational representation of auction-based markets focus on their implementation. In [7], the authors propose an assertive and modular definition of an auction market by representing the market as a set of rules. These rules tackle at first the how and when to bid and assume a single-agent perspective. There is no general reasoning as the semantics is an operational one. The language proposed in [10] adopts an assertive perspective: the proposed language allows the representation of a general auction market, but it is too poor for enabling reasoning. In [1], the authors show how a specific auction, namely combinatorial auctions, can be encoded in a logic program. A hybrid approach mixing linear programming and logic programming has been proposed in [6]: the authors focus on sealed-bid auctions and show how qualitative reasoning helps to refine the optimal quantitative solutions. The closest contribution is the *Market Specification Language* [12] based on GDL [3]. The proposed language is rich enough for representing an auction through a set of rules and then interpreting an auction-instance with the help of a state-based semantics. However, the main limit is the single-agent perspective.

2 CONTRIBUTION

The current version of ADL focus on single-side auctions: that is one seller and multiple buyers or vice-versa. The language is general enough for taking care of goods' quantity (single or multi-unit) and whether it is open or not (sealed-bid). We focus on the auctioneer perspective: how the auction is organized, how the goods are allocated and how to know if the auction is complete.

ADL is appropriate to represent different types of well-known auctions, such as the English Auction and the Multi-Unit Vickrey Auction. For instance, the rules of an English Auction can be formulated by ADL-formulas as shown Figure 1, where $\mathsf{k}, \mathsf{inc} \in \mathbb{N} \setminus \{0\}$ and $\mathsf{startingBid} \in \mathbb{N}$, represent the amount of bidders, the increment in each bidding turn and the starting bid value, respectively.

In the initial state, no one is bidding and the starting bid value is defined (Rule 1). In each round, the players can accept to raise the bid or decline and thus give up from the auction (Rules 6 and 7). The propositions and variables are updated to the next turn, where the bid

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 1–2, Santiago de Compostela, Spain, August 29-30, 2020.

1. $initial \leftrightarrow first \wedge Bid(\mathsf{startingBid}) \wedge \bigwedge_{r \in N_{eng}} \neg isBidding(r) \wedge \neg currWinner(r)$
2. $terminal \leftrightarrow \neg first \wedge \bigwedge_{r \in N_{eng}} \neg isBidding(r) \vee \bigvee_{r \in N_{eng}} wins(r)$
3. $wins(r) \leftrightarrow (isBidding(r) \vee currWinner(r)) \wedge \bigwedge_{i \neq r \in N_{eng}} \neg isBidding(i)$
4. $payment(r, x) \wedge allocation(r, 1) \leftrightarrow wins(r) \wedge Bid(x)$
5. $payment(r, 0) \wedge allocation(r, 0) \leftrightarrow \neg wins(r)$
6. $legal(accept^r) \leftrightarrow initial \vee isBidding(r)$
7. $legal(decline^r) \leftrightarrow \top$
8. $\bigcirc Bid(add(x, \mathsf{inc})) \leftrightarrow Bid(x) \wedge \bigvee_{r \in N_{eng}} does(accept^r)$
9. $\bigcirc Bid(x) \leftrightarrow Bid(x) \wedge (terminal \vee \bigwedge_{r \in N_{eng}} does(decline^r) \vee \bigvee_{r \in N_{eng}} wins(r))$
10. $\bigcirc isBidding(r) \leftrightarrow \neg does(decline^r) \vee (isBidding(r) \wedge terminal)$
11. $\bigcirc currWinner(r) \leftrightarrow isBidding(r) \wedge \bigwedge_{y \neq r \in N_{eng}} \neg isBidding(y) \vee r \prec y$
12. $\bigcirc \neg first \leftrightarrow \top$

Figure 1. English Auction represented by Σ_{eng}

is raised if at least one bidder accepts it (Rules 8 to 12). If there is only one or none active bidder, the auction ends (Rule 2). The winner is the last bidder to accept or one of the bidders that accepted before if everyone declines in the current bidding turn (Rule 3). The losers do not pay, while the winner pays the highest bid (Rules 4 and 5).

ADL allow us to derive properties about auction protocols, such as its finiteness and the uniqueness of payment and winner. It also enables us to show in an explicit way what should be assumed about the behavior of a rational bidder, i.e. a player with a private value for the good being auctioned and who tries to maximize a payoff function. For instance, we are able to show that a rational bidder would keep bidding in an English Auction, as long as the bidding value is smaller then her private value. While different sorts of auction protocols can be expressed in ADL, the bids are represented through the agents' actions. ADL actions are predefined in the auction signature and may have integer parameters representing amount and quantity. However, we believe that the use of bidding languages to generate the action set would allow agents' to bid over goods combinations (bundles), quantities and preferences in an easier and more flexible way. A bidding language is better fitted to combinatorial auctions and its addition to ADL will be investigated for future work.

3 PERSPECTIVES

My thesis aims at designing a *General Auction Player* (GAP) that can interpret and reason about the rules governing an auction-based market. To allow an agent to switch between different kinds of markets, the first step is to develop a general Auction Description Language (ADL), a logic-based language for representing the rules of an auction market, which will then allow a GAP to reason strategically in different environments. ADL is a lightweight approach that considers as the starting point GDL, a simple and practical logical language.

During my thesis, there are two main tracks I intend to explore. First, from the auctioneer point of view, my goal is to go explore two main variants of auctions: double-side auction and combinatorial auction. Clearly, ADL is well suited for both of them but requires some extension. Multiple sorts of goods are not yet possible for instance. I aim to define a new language with its characterization focusing on bids (goods, quantity, bundles and preferences).

Second, I want to investigate how ADL-based players may be implemented so that they can reason about the properties of an auction such as the strategy-proof aspect. The key difference, when the player perspective is considered, is the epistemic and strategic aspects: players have to reason about other players' behavior. My approach will adopt a perspective imported from Strategic Reasoning: considering a set of rules describing a market, the agent should be able to "understand" the rules (e.g., how the winner is determined), to reason about her own private information (e.g., what price should she propose) and also about other players' private information (e.g., what she believes about other player's private valuation). The epistemic component will be based in the epistemic extensions of GDL (GDL-III [11] and Epistemic GDL [4]). In order to reduce the model-checking complexity of these approaches, I first aim to explore a conservative extension of GDL, where the belief and knowledge operators are restricted to numerical variables representing private values, thus logical connectives such as disjunction will be avoided.

ACKNOWLEDGEMENTS

This research is supported by the ANR project AGAPE ANR-18-CE23-0013. I am grateful to my tutor and mentor Laurent Perrussel. Our paper describing ADL was accepted at ECAI 2020 [8].

REFERENCES

[1] C. Baral and C. Uyan, 'Declarative specification and solution of combinatorial auctions using logic programming', in *Proc. of Logic Programming and Non-monotonic Reasoning*, pp. 186–199, Berlin Heidelberg, (2001). Springer.
[2] Yann Chevaleyre, Paul E Dunne, Ulle Endriss, Jerome Lang, Michel Lemaitre, Nicolas Maudet, Julian Padget, Steve Phelps, Juan a Rodrguez-Aguilar, and Paulo Sousa, 'Issues in Multiagent Resource Allocation', *Informatica*, **30**(1), 3–31, (2006).
[3] Michael Genesereth and Michael Thielscher, *General game playing*, Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2014.
[4] Guifei Jiang, Dongmo Zhang, Laurent Perrussel, and Heng Zhang, 'Epistemic GDL: A logic for representing and reasoning about imperfect information games', in *IJCAI*, (2016).
[5] Paul Klemperer, 'Auction Theory: A Guide to the Literature', *Journal of Economic Surveys*, (1999).
[6] H. Geun Lee and R. Lee, 'A hybrid approach of linear programming and logic modeling for the market core of sealed bid auctions', *Annals of Operations Research*, **75**, (1997).
[7] K. M. Lochner and M. P. Wellman, 'Rule-based specification of auction mechanisms', in *Proc. of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, (2004).
[8] Munyque Mittelmann and Laurent Perrussel, 'Auction description language (ADL): a general framework for representing auction-based markets', in *ECAI 2020*, ed., G. de Giacomo, Santiago de Compostela, (2020). IOS Press.
[9] Munyque Mittelmann and Laurent Perrussel, 'Game description logic with integers: A GDL numerical extension', in *FoIKS 2020*, eds., Andreas Herzig and Juha Kontinen, Cham, (2020). Springer International Publishing.
[10] Daniel Rolli, Stefan Luckner, Henner Gimpel, and Christof Weinhardt, 'A Descriptive Auction Language', *Electronic Markets*, **16**(1), 51–62, (2006).
[11] Michael Thielscher, 'GDL-III: A proposal to extend the game description language to general epistemic games', in *Proc. of the European Conference on Artificial Intelligence (ECAI)*, volume 285, pp. 1630–1631, Hague, (2016).
[12] Michael Thielscher and Dongmo Zhang, *From General Game Descriptions to a Market Specification Language for General Trading Agents*, 259–274, Springer Berlin Heidelberg, 2010.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 1–2, Santiago de Compostela, Spain, August 29-30, 2020.

Online Probabilistic Interval-based Event Calculus

Periklis Mantenoglou[1,2]

Abstract. Activity recognition systems detect temporal combinations of 'low-level' activities on sensor data. These systems exhibit various types of uncertainty, often leading to erroneous detection. We present an extension of an interval-based activity recognition system which operates on top of a probabilistic Event Calculus implementation. Our proposed system performs online recognition, as opposed to batch processing, thus supporting data streams.

1 MOTIVATION

Event recognition systems process streams of sensor data and infer composite events of interest by means of pattern matching. Commonly, the sensor data contained in the input streams are called 'low-level events', while the instances of the detected patterns in the output are called 'complex events'. In human activity recognition, sensors may provide the system with the coordinates of a person and their orientation, while visual information processing tools may, subsequently, derive *short-term* activities (STAs), which include a person 'walking', 'running', etc. In this case, an event recognition system derives composite *long-term* activities (LTA) such as two people 'moving together', 'meeting' or 'fighting'.

Uncertainty is inherent in activity recognition applications. An input stream may contain STAs expressed as $Pr :: STA$. Pr corresponds to the probability value of the STA which serves as its confidence estimate. A probabilistic event recognition system consumes such streams and derives a collection of complex events with attached probability values.

Point-based event recognition systems compute complex event occurrences at each separate time-point of the stream. Instantaneous event detection, in conjunction with uncertainty in the input, often result in erroneous LTA recognition. Common cases include LTA probability fluctuations and delayed probability escalation for abrupt LTA occurrences. To tackle these issues, an interval-based system may consume the derived LTA probabilities to compute temporal intervals of LTA occurrences and provide more robust recognition.

The goal of this work is to provide an extension of an interval-based event recognition system in order to perform efficient online recognition. PIEC [1] works on top of probabilistic point-based systems and computes intervals of LTA by means of batch processing. We have presented online PIEC (oPIEC), which extends PIEC by supporting data streams [6].

[1] Institute of Informatics & Telecommunications, NCSR Demokritos, Greece, email: pmantenoglou@iit.demokritos.gr

[2] Department of Informatics & Telecommunications, National and Kapodistrian University of Athens, Greece

Video Presentation: `https://youtu.be/Tw7wUTSl5y0`

2 RESEARCH GOALS

A recent survey [3] outlines the main requirements of a stream reasoning system which performs event recognition. In this regard, oPIEC improves upon the underlying point-based system to handle: 1. large data volume by means of online, out-of-core processing and 2. noise and 3. data incompleteness through interval-based recognition.

Contemporary event recognition system specifications are a trade-off between the requirements of [3]. As an example, RTEC [2] is a run-time event recognition system which can integrate complex domain models and handle event streams by incorporating a windowing mechanism along with caching and indexing techniques. It is not designed, however, to cope with noise and uncertainty.

The goal of our research is to enhance the capabilities of point-based event recognition systems in this manner and potentially meet the standards of state-of-the-art stream reasoning systems.

3 COMPONENTS OF THE SYSTEM

The input of oPIEC is a stream of probability values which refer to the occurrence of an LTA at each time-point. Hence, we will firstly discuss the point-based system deriving these probabilities. A brief overview of the batch interval-based system PIEC will follow. Finally, we will discuss our contribution for supporting online recognition.

The **Event Calculus** is a formalism for representing and reasoning about events and their effects [5]. The dialect of the Event Calculus we adopted comprises events, 'fluents' and a linear time model of integer time-points. Fluents are properties which have different values at different points in time. These value are affected by event occurrences. To integrate the Event Calculus into an activity recognition system, we match events with STAs and fluents with LTAs.

Prob-EC is a system based on a probabilistic logic programming implementation of the Event Calculus [8]. Prob-EC handles uncertainty by means of the logic programming framework ProbLog [4]. Since Prob-EC has been proven effective in human activity recognition experiments [8], we used oPIEC on top of Prob-EC for the majority of our experiments.

PIEC (or Probabilistic interval-based Event Calculus) consumes the output of a point-based system, e.g. Prob-EC, and computes the 'probabilistic maximal intervals' (PMIs) of LTAs, i.e. the maximal intervals during which an LTA is said to take place, with a probability above a given threshold. The probability of an interval of some LTA is equal to the average of the LTA probabilities at the time-points that it contains. We define a PMI as follows:

Definition 1 *A **probabilistic maximal interval** $I_{LTA}=[i,j]$ of LTA is an interval such that, given some threshold $\mathcal{T} \in [0,1]$,*

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 3–4, Santiago de Compostela, Spain, August 29-30, 2020.

$P(I_{LTA}) \geq \mathcal{T}$, *and there is no other interval* I'_{LTA} *such that* $P(I'_{LTA}) \geq \mathcal{T}$ *and* I_{LTA} *is a sub-interval of* I'_{LTA}.

Given a dataset of n instantaneous LTA probabilities $In[1..n]$ and a threshold $\mathcal{T}$, PIEC infers all PMIs of that LTA in linear-time. To achieve this, PIEC constructs:

- The $prefix[1..n]$ list containing the cumulative or prefix sums: $\forall\, i \in [1, n]$, $prefix[i] = \sum_{j=1}^{i} In[j] - \mathcal{T}$.
- The $dp[1..n]$ list, which is calculated by traversing the $prefix$ list in reverse order and is defined as: $\forall\, i \in [1, n]$, $dp[i] = \max_{i \leq j \leq n}(prefix[j])$.

PIEC processes a dataset sequentially using two pointers, s and e, indicating, respectively, the starting point and ending point of a potential PMI. Furthermore, PIEC uses the following variable:

$$dprange[s, e] = \begin{cases} dp[e] - prefix[s-1] & \text{if } s > 1 \\ dp[e] & \text{if } s = 1 \end{cases} \quad (1)$$

It can be proven that:

$$dprange[s, e] \geq 0 \Rightarrow \exists e^* : e^* \geq e \ , P([s, e^*]) \geq 0. \quad (2)$$

Consequently, $[s, e^*]$ is a potential PMI. In this case, PIEC increments the e pointer until $dprange$ becomes negative. When $dprange$ becomes negative, PIEC produces the following PMI: $[s, e-1]$. Once a PMI is computed, PIEC increments the s pointer and recalculates $dprange$. By repeating this process, PIEC computes all PMIs of a given dataset.

In contrast to PIEC, **oPIEC** works on data batches $In[i..j]$, with $i \leq j$, which it discards after processing. To ensure correct PMI computation in the absence of the data prior to i, oPIEC identifies the minimal set of time-points that need to be cached in memory. These time-points are cached in the '*support set*' and express the starting points of potential PMIs, i.e., PMIs that may end in the future.

The support set comprises a set of tuples of the form $(t, prefix[t-1])$, where t is a time-point and $prefix[t-1]$ expresses the $prefix$ value of the previous time-point. Since PMI computation involves calculating $dprange[s, e]$, we cache the value $prefix[s-1]$ for each potential starting point – see equation 1.

It has been been proven that a time-point t may be the starting point of a PMI iff:

$$\forall t_{prev} \in [1, t),\ prefix[t_{prev} - 1] > prefix[t-1] \quad (3)$$

Therefore, oPIEC caches the time-points with the currently minimal $prefix[t-1]$ value, and no other time-points.

When processing the data batch $In[i..j]$, oPIEC computes the PMI $[s, e]$ as follows:

- If the PMI is inside the current data batch, i.e. $s \geq i$, it is computed by means of PIEC.
- Otherwise, when $s < i$, the PMI is computed using a variation of PIEC in which the s pointer iterates over the time-points in the support set. Now, the algorithm runs in linear-time with respect to the size of the data batch and the size of the support set, hinting that, to support efficient reasoning in streaming environments, the support set needs to be bound.

oPIEC$^{\mathbf{b}}$ integrates the functions of oPIEC with an algorithm for support set maintenance, i.e. which elements of the support set should be deleted, in order to make room for new ones. Suppose that a time-point t of the current data batch satisfies condition 3 when the support set is full. In brief, oPIEC$^{\mathrm{b}}$ collects every element of the support set along with every candidate element in the set S. Then, it calculates, for every element of S, its 'score range', an interval of real numbers defined as:

$$score\_range[t] = [prefix[t-1], prefix[prev_s[t] - 1]), \quad (4)$$

where $prev_s[t]$ is the time-point before t in S. The longer the $score\_range[t]$, i.e. the longer the distance between $prefix[t-1]$ and $prefix[prev_s[t] - 1]$, the more likely it is, intuitively, that a future time-point t_e will arrive with $dp[t_e] \in score\_range[t]$, and thus that t will be the starting point of a PMI. Hence, oPIEC$^{\mathrm{b}}$ removes the elements with shortest score range from the support set.

4 CONCLUSION AND FUTURE WORK

oPIEC$^{\mathrm{b}}$ has been thoroughly tested on a benchmark dataset for human activity recognition. To generate the input stream of instantaneous probabilities, we used Prob-EC, as well as OSLα [7]. These experiments proved the efficacy of our system. oPIEC$^{\mathrm{b}}$ reaches the predictive accuracy of PIEC with a small support set, as compared to the size of the stream.

Since the accuracy of oPIEC$^{\mathrm{b}}$ depends on the correctness of the support set, we aim at developing new support set maintenance techniques, and thus reduucing the memory requirements of our system.

To further investigate on the predictive accuracy of oPIEC, we intent to compare it with other state-of-the-art event recognition systems, as well as target other domain models, apart from human activity recognition, like the maritime domain.

Finally, we plan to revise the definition of a 'probabilistic maximal interval' in an attempt to optimise the recognition of our system with regards to the uncertainty of the underlying point-based system.

ACKNOWLEDGEMENTS

This work was supported by the INFORE project, which has received funding from the European Union's Horizon 2020 research and innovation programme, under grant agreement No 825070.

REFERENCES

[1] A. Artikis, E. Makris, and G. Paliouras, 'A probabilistic interval-based event calculus for activity recognition', *Annals of Mathematics and Artificial Intelligence*, (Aug 2019). https://doi.org/10.1007/s10472-019-09664-4.

[2] A. Artikis, M. Sergot, and G. Paliouras, 'An event calculus for event recognition', *IEEE Transactions on Knowledge and Data Engineering*, **27**(4), 895–908, (2015).

[3] D. Dell'Aglio, E. Della Valle, F. van Harmelen, and A. Bernstein, 'Stream reasoning: A survey and outlook', *Data Sci.*, **1**, 59–83, (2017).

[4] A. Kimmig, B. Demoen, L. De Raedt, V. Santos Costa, and R. Rocha, 'On the implementation of the probabilistic logic programming language ProbLog', *TPLP*, **11**, 235–262, (2011).

[5] R. Kowalski and M. Sergot, 'A logic-based calculus of events', *New Generation Computing*, **4**(1), 67–95, (1986).

[6] P. Mantenoglou, A. Artikis, and G. Paliouras, 'Online probabilistic interval-based event calculus', in *Proceedings of ECAI*, (2020). http://cer.iit.demokritos.gr/publications/papers/2020/ecai2020.pdf.

[7] E. Michelioudakis, A. Skarlatidis, G. Paliouras, and A. Artikis, 'Oslα: Online structure learning using background knowledge axiomatization', in *Proceedings of ECML-PKDD*, (2016).

[8] A. Skarlatidis, A. Artikis, J. Filippou, and G. Paliouras, 'A probabilistic logic programming event calculus', *Theory and Practice of Logic Programming*, (2013).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 3–4, Santiago de Compostela, Spain, August 29-30, 2020.

Learning from the Individuals and the Crowd in a Society of Devices

Fernando E. Casado[1]

Abstract. State-of-the-art machine learning methods have several drawbacks when facing real distributed problems, in which the available information is biased, evolves over time, and is irregularly and heterogeneously distributed among the different devices. We propose a new approach for learning in society, called *glocal* learning, where devices get certain prominence back, having to learn locally, continuously, and autonomously, but also improving models globally, in the cloud, combining what each device has learned locally.

1 CONTEXT AND MOTIVATION

Nowadays, we live in a context in which smart devices are almost ubiquitous. Smartphones, tablets, wearables, robots and "things" from the Internet of Things (IoT) are already counted in millions and allow a growing and sophisticated number of applications related to absolutely all human domains: education, health, leisure, travel, banking, sport, social interaction, etc. Most of these devices are already interconnected or connected to the cloud and can collect huge amounts of data from the environment thanks to the fact that they are becoming more and more sensorized. Devices can take advantage of this data by learning models in order to adapt and improve their behavior, thus benefiting the consumer. Nevertheless, how to do this effectively and efficiently is an open question.

Traditional and offline server-based machine learning techniques are the most common approach to carry out the learning process. In the context of distributed devices, this involves collecting data from all the devices. This data, typically sensor measurements, photos, videos, and location information, must be later uploaded and processed centrally in a cloud-based server or data center. Thereafter, the data is used to provide insights or produce effective inference models. With this approach, in the last decade, *deep learning* techniques have proven to be very accurate. However, in this kind of distributed scenarios, cloud-centric learning can be highly inefficient or even infeasible, since it involves facing the following challenges:

- **Scalability**, in terms of storage, communication, and computational costs. Central storage is never a scalable solution. Transferring large amounts of data over the network can take a long time, and communication may be a continuous overhead. Similarly, the computational cost of central processing is much bigger than the sum of the costs of parallel computing.
- **Data privacy and sensitivity**. Central data collection puts user privacy into risk. In recent years, governments have been implementing data privacy legislations in order to protect the consumer, limiting data collection and storage only to what is consented by the consumer and absolutely necessary for processing.
- **Nonstationarity**. Most machine learning algorithms assume that data comes from a static distribution. However, it does not hold in many real-world applications, where the underlying distribution of the data streams changes over time in unforeseen ways. This phenomenon is widely known as *concept drift* [5]. If a concept drift occurs, the inducted pattern of past data may not be relevant to the new data, leading to poor outcomes. This problem is related to two conflicting objectives: retaining previously learned knowledge that is still relevant and replacing the obsolete one with current information. This is known as the *stability-plasticity dilemma*.
- **Acquisition of relevant labeled data**. Gathering enough relevant labeled data is typically a complex task because it usually involves human intervention. In real-world problems, training data may not have labels or these may be poisoned or mistaken, leading to runtime mispredictions.

[1] CiTIUS, Universidade de Santiago de Compostela, Spain, email: fernando.estevez.casado@usc.es

2 RELATED WORK

Taking into account the aforementioned challenges, there have emerged some adaptations of the cloud-centric paradigm that suggest that the deep learning process can take place in the cloud and the learned model can be then transferred to the device, which is where it is employed [7]. Other hybrid proposals suggest that there may be a global training in the cloud and a final adjustment of the model at the local level, or just the other way round, a pre-training of the model in the device and a final training in the cloud [4]. However, any of these strategies would still involve moving a significant volume of potentially sensitive data to perform the cloud stage.

A better option for learning in this kind of scenarios where data is naturally distributed seems to be *distributed learning* [8]. Different from the classical cloud-centric approach, in distributed learning the learning process can be carried out in a distributed and parallel manner along a network of interconnected devices that are able to collect, store, and process their own data locally. Once each device performs its local learning, a global integration stage is typically carried out in the cloud, so that a global model is agreed upon from local models. There have been proposed many powerful distributed optimization algorithms to solve the local sub-problems. Many of these proposals are based on gradient descent or other optimization algorithms, such as augmented Lagrangian methods, being *Alternating Direction Method of Multipliers* (ADMM) the most popular one [1]. There are also some approaches based on ensemble learning, such as *Effective Stacking* [9].

In recent years, a new framework called *federated learning* and its most widely known algorithm, *Federated Averaging* (FedAvg), have been boosted by Google [6]. The core idea of federated learning is similar to that of distributed learning., i.e., to solve local problems on

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 5–6, Santiago de Compostela, Spain, August 29-30, 2020.

the devices and aggregate updates on a server without uploading the original user data. However, the goals pursued by each are different. Distributed learning mainly focus on parallelizing computing power, and usually assumes that the local datasets are identically distributed and roughly have the same size. None of these hypotheses are made for federated learning. Instead, datasets are typically heterogeneous and their sizes may span several orders of magnitude. Another key area of attention for federated learning is data privacy.

Allocating the learning process among the network of devices is a natural way of scaling up learning algorithms. Furthermore, it makes it easier to protect users' privacy, since sharing raw data with the cloud can be avoided. Therefore, distributed and federated learning are much more suitable for the context of devices than traditional server-based learning. Nevertheless, there are still some issues that these paradigms do not address or only partially address. The most important one is the aforementioned nonstationarity. In order to deal with this problem, there is another framework called *continual learning* [5]. Continual learning is built on the idea of learning continuously and adaptively about the external world, being able to smoothly update the prediction model to take into account different tasks and data distributions but still being able to re-use and retain useful knowledge during time. Hence, continual learning is the only paradigm which forces us to deal with a higher and realistic timescale where data becomes available only during time, we have no access to previous perception data and it is imperative to build on top of previously learned knowledge. However, continual learning does not address how to learn in distributed contexts.

3 OUR PROPOSAL

Each learning proposal mentioned so far is focused on providing a solution to a specific challenge, but normally neglects others. For example, federated learning focuses on scalability and user privacy, but it does not cope with nonstationarity. This is addressed by continual learning, which, however, does not take into account issues such as learning from distributed data or user privacy. Therefore, we have developed a new paradigm of learning in society, called *glocal learning* [2, 3], intending to address all the aforementioned challenges at the same time. Our approach consists of achieving local models, on the devices themselves, but which are later agreed upon globally, in the cloud. This global model is then returned to the devices so that they speed up the local learning. This process of global consensus and local adaptation can be repeated indefinitely over time.

At the local level, the device processes the data that it captures from its environment. Local learning gains more prominence than in distributed or federated proposals. There is a semi-supervised transduction module to make the most of unlabeled data in semi-supervised contexts. Autonomously, the device will be able to make decisions about what data to store and for how long, when to train a model, and when to update it. This is possible because our proposal explicitly addresses the detection of concept drifts.

In the cloud, a global model is agreed upon from the local models that are shared by the devices. Different from a simple average of weights as it is usually done in federated learning, our consensus system includes a voting algorithm, which we call Distributed Effective Voting [3]. Using this method, the participating devices are evaluated against each other, which help us decide which local models to rely on and which are dispensable in global integration.

The fact that only models are moved amongst the local devices and the cloud, helps to reduce communication costs and protect local privacy. Learning in a distributed and parallel way, being robust to concept drift, also improves scalability in terms of storage and computing.

4 RESULTS AND FUTURE WORK

So far, we have limited our proposal to supervised and semi-supervised contexts, and in particular for classification tasks. We have applied it in a heterogeneous community of smartphone users to solve binary and multiclass classification tasks related to human activity recognition. The results [3] show the advantages that *glocal* learning provides with respect to other state-of-the-art approaches.

We believe that we have opened a very promising line of research and a large amount of work can still be done. To date, we have proposed a general architecture and a first implementation of each of its components [3]: local learning, semi-supervised labeling, drift detection and adaptation, and selection of local candidates for the subsequent global consensus. However, we believe that all of these parts have a large scope for improvement. For example, we want to explore optimal ways to update local models and to combine them in the cloud. We also want to include new components, such as *glocal* feature and instance selection, and to increase its applicability.

ACKNOWLEDGEMENTS

This research was supported by AEI/FEDER (grant number TIN2017-90135-R), as well as *Consellería de Educación, Universidade e Fromación Profesional* of Galicia (accreditation ED431G/01, ED431G/08, and ED431C2018/29), and the *Ministerio de Universidades* of Spain (FPU17/04154). Special thanks to my supervisors, Roberto Iglesias and Senén Barro, as well as my colleagues Carlos V. Regueiro and Dylan Lema, for their invaluable support.

REFERENCES

[1] Fredrik Andersson, Marcus Carlsson, Jean-Yves Tourneret, and Herwig Wendt, 'A new frequency estimation method for equally and unequally spaced data', *IEEE Transactions on Signal Processing*, **62**(21), 5761–5774, (2014).

[2] Fernando E Casado, Dylan Lema, Roberto Iglesias, Carlos V Regueiro, and Senén Barro, 'Learning from the individuals and the crowd in robotics and mobile devices', in *Iberian Robotics conference*, pp. 632–643. Springer, (2019).

[3] Fernando E Casado, Dylan Lema, Roberto Iglesias, Carlos V. Regueiro, and Senén Barro, 'Collaborative and continual learning for classification tasks in a society of devices', *arXiv preprint arXiv:2006.07129*, (2020).

[4] Nicholas D Lane and Petko Georgiev, 'Can deep learning revolutionize mobile sensing?', in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, pp. 117–122. ACM, (2015).

[5] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez, 'Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges', *Information Fusion*, **58**, 52–68, (2020).

[6] H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Aguera y Arcas, 'Federated learning of deep networks using model averaging', *arXiv preprint arXiv:1602.05629v1*, (2016).

[7] Preetum Nakkiran, Raziel Alvarez, Rohit Prabhavalkar, and Carolina Parada, 'Compressing deep neural networks using a rank-constrained topology', in *Proceedings of the 16th Anual Conference of the International Speach Communication Association*, pp. 1473–1477. ISCA, (2015).

[8] Diego Peteiro-Barral and Bertha Guijarro-Berdiñas, 'A survey of methods for distributed machine learning', *Progress in Artificial Intelligence*, **2**(1), 1–11, (2013).

[9] Grigorios Tsoumakas and Ioannis Vlahavas, 'Effective stacking of distributed classifiers', in *Ecai*, volume 2002, pp. 340–344, (2002).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 5–6, Santiago de Compostela, Spain, August 29-30, 2020.

Monitoring cardiovascular risk by facial video processing and fuzzy rules

Gianluca Zaza[1]

Abstract. Remote photoplethysmography has been recently used to monitor vital parameters by avoiding the use of contact devices. The goal of this work is to develop a smart mirror intended as a contactless device that captures video frames of the mirrored face and provides a real-time estimation of vital parameters through remote photopletismography. The mirror-based system embeds an AI component based on fuzzy inference rules to predict the level of cardiovascular risk using the estimated parameters as input.

1 INTRODUCTION

Cardiovascular disease is one of the main death causes in the world according to a 2015 report[2]. To prevent the onset of chronic diseases and detect the risk level of cardiovascular diseases [6], it is essential to monitor vital parameters such as heart rate (HR), respiratory rate (BR) and blood oxygen saturation (SpO_2). Conventional methods for measuring cardiovascular parameters use skin contact techniques requiring wearable devices (ECG). Although non-invasive and low cost optical devices (e.g. pulse oximeter), using photoplethysmography (PPG) can detect the cardiovascular pulse wave through variations in transmitted or reflected light [1], they still require direct contact to the skin. For this reason, these devices are unconfortable. An alternative approach is to use a video processing algorithms extract a remote photoplethysmographic (rPPG) signal [17]. Then, through the use of image processing methods and blind source separation algorithms, vital parameters are estimated [12]. This paper provides a synthetic description of the ongoing PhD research activity of the author. A more detailed description can be found in [10, 11, 3, 2, 4]. We propose a non-contact device for monitoring vital parameters based on a see-through mirror with a camera capturing facial video frames of patients. Through the processing of the rPPG signal extracted from the video frames, HR, BR and SpO_2 are measured. Also lip color is automatically estimated by using clustering-based color quantization. Finally, a hierarchical fuzzy system (HFIS) is integrated in the mirror-based monitoring system to infer the cardiovascular risk level. The goal of the research is to develop a pervasive device that can be easily used both for personal monitoring in domestic environments and for supporting medical diagnosis through telemedicine.

2 MATERIALS AND METHODS

The proposed system estimates vital parameters in real-time through the use of a see-through mirror made of an acrylic film equipped with a camera that acquires the facial video frames useful to extract the photopletismographic signal arising from facial blood vessels. Good light is guaranteed by a wood frame including two strips of white LED lights on the right and left side of the mirror. In the current prototypical version of the system, the computer is an All-in-One equipped with CPU Intel(R) Core(TM) i5-5200 2.20GHz 64 bit, 4GB RAM and 500GB hard disk.

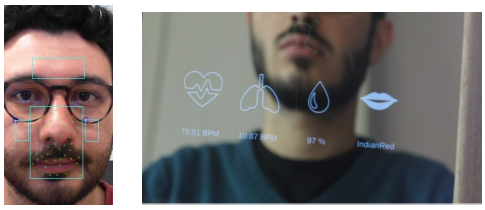

Figure 1. Location of ROIs and tracking points (left figure) and result of vital parameter estimation through the smart mirror (right figure).

The first step is the real-time identification and acquisition of the face through the camera. A preliminary 26-seconds video frames cycle is acquired to avoid camera distortion; the following frames are acquired every 2 seconds. To capture and process video frames we used the Python *OpenCV*[3] library; to detect the face inside video frames we used a pretrained frontal detector, available with the *Dlib*[4] library. Then, we apply a *Facial Landmark Detection*[5] to obtain a set of 68 facial landmarks [9]. Next we identified three regions of interest (ROI) with a strong passage of blood modulation, namely a forehead ROI (90×30 px) and two (right and left) cheek ROIs (20×30 px). To reduce signal distortion due to facial movements, we applied the Kanade-Lukas-Tomasi tracking method [13, 16]. We added a new ROI (75×80 px) in the center of the face to join the 3 ROIs and build a motion matrix that helps keeping track of the ROI coordinates between the subsequent frames. Fig. 1(left) shows the ROIs and tracking points on the face.

To estimate the vital parameters we analyze the RGB signals coming from the ROIs. Each ROI is separated into RGB channels and then for each channel all ROIs pixels are averaged to obtain a V_{RGB} signal matrix. Afterwards, the Finite Impulse Response [15] and the Chrominance method [7] are used to eliminate high frequency noise in signals and obtain a more robust signal to lighting variations. To ensure uniform signal sampling we apply the linear interpolation method. Finally, to discover the most informative signal for the evaluation of vital signals [12, 17, 8] we use the Power spectral density

[1] Computer Science Department, University of Bari Aldo Moro, Italy, email: gianluca.zaza@uniba.it

[2] www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html (accessed on 2 March 2018).

[3] OpenCV: *https://pypi.org/project/opencv-python/*

[4] Dlib: *http://dlib.net/*

[5] Facial Landmark Detection: *http://dlib.net/imaging.html#shape_predictor*

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 7–8, Santiago de Compostela, Spain, August 29-30, 2020.

through Welch's method [14]. As an additional vital parameter, the color of lips is obtained through the extraction and analysis of the specific ROI including the mouth. K-means clustering is applied to quantify the predominant color in the ROI, thus obtaining the 3 main colours in RGB format. Finally, we converted the color from RGB to hexadecimal format and then from hexadecimal format to a linguistic term using the *Name-That-Color* library[6]. Fig. 1(right) shows the output of the mirror-based system after estimation of vital parameters.

Besides measurement of vital parameters, the system includes a fuzzy diagnosis component to predict the risk of cardiovascular disease. To this aim we have developed a Hierarchical Fuzzy Inference System (HFIS) starting from a flat fuzzy rule base previously defined with the help of a physician [2]. We created the linguistic variables and related fuzzy sets using the FISDET tool [5]. The HFIS consists of 3 FIS organized in a hierarchical fashion. Each FIS has 2 input variables and 1 output variable. We created intermediate input/output variables between levels named Y_1 and Y_2 with their respective language terms. The last level provides the final output which represents the level of cardiovascular risk. On the overall, 27 rules have been defined from the combination of input and output variables present in each HFIS level.

3 PRELIMINARY RESULTS

We conducted a preliminary experiment to evaluate the goodness of measurements made through the mirror-based system by comparing it to a standard pulse oximeter. We considered 21 people (4 women and 17 men) aged between 22 and 37 years. During the tests we asked each user to stand in front of the mirror at a distance of about 50 cm. We found that measurements obtained by the mirror have a good correlation with measurements obtained by the pulse oximeter.

Then we tested the accuracy of the HFIS by comparing it with the flat FIS created in the previous work [2]. The testing data were collected by asking experts to evaluate and label a dataset of 116 subjects containing vital signs and assign each label a risk level (Low, Medium, High, Very High). We created 12 HFIS configurations, considering all combinations of input variables. The best HFIS was compared with the flat FIS. The overall accuracy on testing data was 71.55% for the best HFIS and 69.97% for the flat FIS. The HFIS has also a smaller rule base (27 fuzzy rules) compared to the flat FIS (81 rules).

Table 1 compares results for the two fuzzy systems. It can be seen that TNR (True Negative Rate) and NPV (Negative Predict Value) are generally higher than TPR (True Positive Rate) and PPV (Positive Predict Value) for the HFIS. This means that the knowledge base is more effective in determining the non-belongingness to each class than the sensitivity to each specific risk level. This is probably due to the imbalance of the data set. HFIS better discriminates the extreme risk classes (Low and Very High), while the original FIS better recognizes intermediate risk classes (Medium and High).

Table 1. Classification results of HFIS and FIS.

	HFIS					FIS				
Risk	ACC	TNR	TPR	PPV	NPV	ACC	TNR	TPR	PPV	NPV
Low	0,76	0.80	0,14	0.04	0.93	0.83	1.00	0.77	1.00	0.60
Medium	0,76	0.80	0.14	0.04	0.93	0.75	0.76	0.57	0.13	0.96
High	0.91	0.97	0.12	0.25	0.94	0.91	0.94	0.50	0.40	0.96
Very high	0.91	0.97	0.53	0.73	0.93	0.88	0.96	0.40	0.60	0.91

[6] http://chir.ag/projects/name-that-color/

4 CONCLUSION AND FUTURE WORKS

This work aims to create a low-cost, non-invasive and easy-to-use solution to monitor vital parameters and predict a risk of cardiovascular disease. Preliminary experiments have shown that our device provides accurate measurement of vital parameters, as well as reliable risk prediction thanks to the intelligent component based on fuzzy rules. Future works aim to conduct large-scale experiments in hospitals in order to acquire more data from patients suffering from cardiovascular diseases. Furthermore, we intend to use machine learning techniques to automatically extract fuzzy rules from data.

ACKNOWLEDGEMENTS

The author wishes to thank Prof. Giovanna Castellano for supervising his PhD research activity.

REFERENCES

[1] J. Allen, 'Photoplethysmography and its application in clinical physiological measurement', *Physiological measurement*, **28**(3), R1, (2007).

[2] G. Casalino, G. Castellano, C. Castiello, V. Pasquadibisceglie, and G. Zaza, 'A fuzzy rule-based decision support system for cardiovascular risk assessment', in *Fuzzy Logic and Applications*, eds., R. Fullér, S. Giove, and F. Masulli, pp. 97–108, (2019).

[3] G. Casalino, G. Castellano, V. Pasquadibisceglie, and G. Zaza, 'Contact-less real-time monitoring of cardiovascular risk using video imaging and fuzzy inference rules', *Information*, **10**, 9, (2018).

[4] G. Casalino, G. Castellano, V. Pasquadibisceglie, and G. Zaza, 'Evaluating end-user perception towards a cardiac self-care monitoring process', in *Wireless Mobile Communication and Healthcare*, ed., Gregory M.P. et al. O'Hare, pp. 43–59, (2020).

[5] G. Castellano, C. Castiello, V. Pasquadibisceglie, and G. Zaza, 'Fisdet: Fuzzy inference system development tool', *Int. J. Comput. Intell. Syst.*, **10**, 13–22, (2017).

[6] S. Cook, M. Togni, M. C. Schaub, P. Wenaweser, and O. M. Hess, 'High heart rate: a cardiovascular risk factor?', *European heart journal*, **27**(20), 2387–2393, (2006).

[7] Gerard De Haan and Vincent Jeanne, 'Robust pulse rate from chrominance-based rppg', *IEEE Transactions on Biomedical Engineering*, **60**(10), 2878–2886, (2013).

[8] L. Kong et al., 'Non-contact detection of oxygen saturation based on visible light imaging device using ambient light', *Optics express*, **21 15**, 17464–71, (2013).

[9] V. Kazemi and J. Sullivan, 'One millisecond face alignment with an ensemble of regression trees', *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874, (2014).

[10] V. Pasquadibisceglie, G. Zaza, and G. Castellano., 'A personal healthcare system for contact-less estimation of cardiovascular parameters', in *Proc. of the AEIT International Annual Conference (AEIT2018)*, pp. 1–6, Bari, Italy, (October 3-5, 2018). IEEE.

[11] V. Pasquadibisceglie, G. Zaza, and G. Castellano, 'A personal healthcare system for contact-less estimation of cardiovascular parameters', *2018 AEIT International Annual Conference*, 1–6, (2018).

[12] M. Z. Poh, D. J. McDuff, and R. W. Picard, 'Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.', *Optics express*, **18**(10), 10762–10774, (2010).

[13] J. Shi and C. Tomasi, 'Good features to track', in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600, (1994).

[14] OM Solomon Jr, 'Psd computations using welch's method', *STIN*, **92**, 23584, (1991).

[15] T Speake and R Mersereau, 'A note on the use of windows for two-dimensional fir filter design', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **29**(1), 125–127, (1981).

[16] C. Tomasi and T. Kanade, 'Detection and tracking of point features', in *Technical Report CMU-CS-91-132*, School of Computer Science, Carnegie Mellon Univ. Pittsburgh, (1991).

[17] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, 'Remote plethysmographic imaging using ambient light.', *Optics express*, **16**(26), 21434–21445, (2008).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 7–8, Santiago de Compostela, Spain, August 29-30, 2020.

Towards a Ranking-Based Semantics using System Z

Kenneth Skiba [1]

Abstract. We discuss ranking arguments from an Dung-style argumentation framework with the help of conditional logics. Using an intuitive translation for an argumentation framework to generate conditionals, we can apply nonmonotonic inference systems to generate a ranking on these conditionals. With this ranking we construct a ranking for our arguments.

1 INTRODUCTION

Formal argumentation [2] describes a family of approaches to modeling rational decision-making through the representation of arguments and their relationships. A particular important representative approach is that of abstract argumentation [6], which focuses on the representation of arguments and a conflict relation between arguments through modeling this setting as a directed graph. Here, arguments are identified by vertices and an *attack* from one argument to another is represented as a directed edge. This simple model already provides an interesting object of study, see [3] for an overview. Reasoning is usually performed in abstract argumentation by considering *extensions*, i. e., sets of arguments that are jointly acceptable given some formal account of "acceptability". Therefore, this classical approach differentiates between "acceptable" arguments and "rejected" arguments.

However, ranking-based semantics [1] provide a finer-grained assessment of arguments. For this idea there is already a line of work like a ranking with respect to a categoriser function [11] or based on a two-person zero-sum strategic game [10] and many more. [5] summarizes the state-of-the-art models for ranking arguments.

We want to make some first steps towards the use of conditional logic and the System Z inference mechanism [7] to define rankings between arguments. Conditional logic is a general non-monotonic representation formalism that focuses on default rule of the form "if A then B" and there exist some interesting relationships between this formalism and that of formal argumentation [8, 9]. We make use of these relationships here for the purpose of defining a novel ranking-based semantics for abstract argumentation.

2 ABSTRACT ARGUMENTATION FRAMEWORKS

In this work, we use *argumentation frameworks* first introduced in [6]. An *argumentation framework* AF is a pair $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is a finite set of arguments and $\mathcal{R}$ is a set of attacks between arguments with $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. An argument a is said to *attack* b if $(a, b) \in \mathcal{R}$. We call an argument a *acceptable with respect to a set* $S \subseteq \mathcal{A}$ if for each attacker $b \in \mathcal{A}$ of this argument a with $(b, a) \in \mathcal{R}$, there is an argument $c \in S$ which attacks b, i. e., $(c, b) \in \mathcal{R}$; we then say that a is *defended by* c. An argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$ can be illustrated by a directed graph with vertex set $\mathcal{A}$ and edge set $\mathcal{R}$.

Example 1. Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ with $\mathcal{A} = \{a, b, c, d\}$ and $\mathcal{R} = \{(a, b), (b, c), (c, d), (d, c)\}$ be an argumentation framework. Argument b is not acceptable with respect to any set S of arguments, as b is not defended against a's attack. On the other hand, c is acceptable with respect to $S = \{a, c\}$, as a defends c against b's attack and c defends itself against d's attack.

Up to this point the arguments can only have the two statuses of accepted or not accepted[2], but we want to have a more fine-graded comparison between arguments. For this we use the idea of ranking-based semantics [1, 5].

Definition 2 (Ranking-based semantics)**.** A *ranking-based semantics* σ associates to any argumentation framework $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ a preorder $\succeq^{\sigma}_{AF}$ on $\mathcal{A}$. $a \succeq^{\sigma}_{AF} b$ means that a is at least as acceptable as b. With $a \simeq^{\sigma}_{AF} b$ we describe that a and b are equally acceptable, i. e., $a \succeq^{\sigma}_{AF} b$ and $b \succeq^{\sigma}_{AF} a$. Finally we say a is strictly more acceptable than b, denoted by $a \succ^{\sigma}_{AF} b$, if $a \succeq^{\sigma}_{AF} b$ and not $b \succeq^{\sigma}_{AF} a$. We denote by $\sigma(AF)$ the ranking on $\mathcal{A}$ returned by σ.

3 CONDITIONAL LOGICS

We use a set of atoms A and connectives $\wedge$ (and), $\vee$ (or), and $\neg$ (negation) to generate the *propositional language* $\mathcal{L}(A)$. w is an *interpretation* (or *possible world*) for $\mathcal{L}(A)$ when $w : A \to \{\text{TRUE}, \text{FALSE}\}$. We denote the set of all interpretations as $\Omega(A)$. An interpretation w *satisfies* an atom $a \in A$ ($w \vdash a$), if and only if $w(a) = \text{TRUE}$. The relation $\vdash$ is extended to arbitrary formulas in the usual way. We will abbreviate an interpretation w with its *complete conjunction*, i. e., if $a_1, \ldots, a_n \in A$ are the atoms that are assigned TRUE by w and $a_{n+1}, \ldots, a_m \in A$ are the ones assigned with FALSE, w will be identified with $a_1 \ldots a_n \overline{a_{n+1}} \ldots \overline{a_m}$. For $\Phi \subseteq \mathcal{L}\{A\}$ we define $w \vdash \Phi$ if and only if $w \vdash \phi$ for every $\phi \in \Phi$. With $Mod(X) = \{w \in \Omega(A) | w \vdash X\}$ we define the set of models for a set of formulas X. A *conditional* is a structure of the form $(\varphi|\psi)$ and represents a rule "If ψ than (usually) ϕ".

We can consider conditionals as *generalized indicator functions* [4] for possible worlds w as follows:

$$((\varphi|\phi))(w) = \begin{cases} 1 : w \vdash \phi \wedge \varphi \\ 0 : w \vdash \phi \wedge \neg\varphi \\ u : w \vdash \neg\phi \end{cases} \quad (1)$$

where u stands for *unknown*. Informally speaking, a world w *verifies* a conditional $(\varphi|\phi)$ iff it satisfies both antecedent and conclusion

[1] Institute for Web Science and Technologies, University of Koblenz-Landau, Germany, email: kennethskiba@uni-koblenz.de

[2] However, using labeling-based semantics we can generate a three-valued model [14].

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 9–10, Santiago de Compostela, Spain, August 29-30, 2020.

$((\varphi|\phi)(w) = 1)$; it *falsifies* iff is satisfies the antecedence but not the conclusion $((\varphi|\phi)(w) = 0)$; otherwise the conditional is *not applicable* $((\varphi|\phi)(w) = u)$. A conditional $(\varphi|\phi)$ is satisfied by w if it does not falsify it.

Semantics are given to sets of conditionals via ranking functions [7,13]. With a ranking function, also called *ordinal conditional function (OCF)*, $\kappa : \Omega(A) \rightarrow \mathbb{N} \cup \{\infty\}$ we can express the degree of plausibility of possible worlds $\kappa(\phi) := min\{\kappa(w)|w \vdash \phi\}$. With the help of OCFs κ we can express the acceptance of conditionals and nonmonotonic inferences, so $(\varphi|\phi)$ is accepted by κ iff $\kappa(\phi \wedge \varphi) < \kappa(\phi \wedge \varphi)$. With $Bel(\kappa) = \{\phi|\forall w \in \kappa^{-1}(0) : w \vdash \phi\}$ we denote the most plausible worlds.

As there are an infinite number of ranking functions that accept a given set of conditionals, we consider System Z [7] as an inference relation, which yields us a uniquely defined ranking function for reasoning.

Definition 3 (System Z)**.** $(\varphi|\phi)$ is tolerated by a finite set of conditionals Δ if there is a possible world w with $(\phi|\varphi)(w) = 1$ and $(\phi'|\varphi')(w) \neq 0$ for all $(\phi'|\varphi') \in \Delta$. The *Z-partition* $(\Delta_0, \ldots, \Delta_n)$ of Δ is defined as:

- $\Delta_0 = \{\delta \in \Delta | \Delta \text{ tolerates } \delta\}$
- $\Delta_1, \ldots, \Delta_n$ is the Z-partition of $\Delta \setminus \Delta_0$

For $\delta \in \Delta$: $Z_\Delta(\delta) = i$ iff $\delta \in \Delta_i$ and $\Delta_1, \ldots, \Delta_n$ is the Z-partitioning of Δ.

We define a ranking function $\kappa^Z_\Delta : \Omega \rightarrow \mathbb{N} \cup \{\infty\}$ as $\kappa^Z_\Delta(w) = max\{Z(\delta)|\delta(w) = 0, \delta \in \Delta\} + 1$, with $max\ \emptyset = -1$. Finally $\Delta \mid\sim_Z \phi$ if and only if $\phi \in Bel(\kappa^Z_\Delta)$.

Example 4. Let $\Delta = \{(a|\neg b), (b|\neg a), (c|\neg b \wedge \neg a \wedge \neg d), (d|\top), (c|\neg d)\}$. For this set of conditionals, $\Delta = \Delta_0 \cup \Delta_1$ with $\Delta_0 = \{(a|\neg b), (b|\neg a), (c|\neg b \wedge \neg a \wedge \neg d)\}$ and $\Delta_1 = \{(\neg a \wedge \neg b|\ d)\}$ therefore we have the values from the following table.

ω	$Z((a\|\neg b))$	$Z((b\|\neg a))$	$Z((c\|\neg b \wedge \neg a \wedge \neg d))$	$Z((d\|\top))$	$Z((\neg a \wedge \neg b\|d))$
$abcd$	u	u	u	1	0
$abc\bar{d}$	u	u	u	0	u
$ab\bar{c}d$	u	u	u	1	0
$ab\bar{c}\bar{d}$	u	u	u	0	u
$a\bar{b}cd$	1	u	u	1	0
$a\bar{b}c\bar{d}$	1	u	u	0	u
$a\bar{b}\bar{c}d$	1	u	u	1	0
$a\bar{b}\bar{c}\bar{d}$	1	u	u	0	u
$\bar{a}bcd$	u	1	u	1	0
$\bar{a}bc\bar{d}$	u	1	u	0	u
$\bar{a}b\bar{c}d$	u	1	u	1	0
$\bar{a}b\bar{c}\bar{d}$	u	1	u	0	u
$\bar{a}\bar{b}cd$	0	0	u	1	1
$\bar{a}\bar{b}c\bar{d}$	0	0	1	0	u
$\bar{a}\bar{b}\bar{c}d$	0	0	u	1	1
$\bar{a}\bar{b}\bar{c}\bar{d}$	0	0	0	0	u

Table 1. Values for Example 4

So we can derive $(\kappa^Z_{\Delta_0})^{-1}(0) = \{abcd, ab\bar{c}d, a\bar{b}cd, a\bar{b}\bar{c}d, \bar{a}bcd, \bar{a}b\bar{c}d\}$ and $(\kappa^Z_{\Delta_1})^{-1}(0) = \emptyset$.

In some recent works [12] we have already presented these first steps. In these works we first translated an abstract argumentation framework into a conditional logic base, using methods introduced in [8,9]. Using the constructed logic base as well as ranking functions like *System Z* we can then rank the resulting worlds based on their plausibility. So the most plausible world is ranked the highest. Based on this ranking we constructed a way to rank the arguments, which results in an ranking-based semantics [1,5]. The first idea is to count the appearances of each argument in the most plausible worlds. So for *System Z* we count how often a positive literal of an argument is inside $(\kappa^Z_{\Delta_0})^{-1}(0)$.

One approach to evaluate an ranking-based semantics is to check the properties, which this semantics satisfies and which it violates. For that we should first look at the simplest properties like *Void Precedence* [1,10], which says that an argument which is not attack should always be ranked higher then an argument with an attack, or *Self-Contradiction* [10], meaning that an argument which contradict itself should always be ranked worse then any other argument.

This approach for ranking arguments can lead to a new way of reasoning inside formal argumentation and also bridging the gap of plausible conditionals and formal argumentation.

ACKNOWLEDGEMENTS

The research reported here was supported by the Deutsche Forschungsgemeinschaft under grant KE 1413/11-1.

I have a full paper accepted at the ECAI2020 conference: K.Skiba, D.Neugebauer and J.Rothe. 'Complexity of Possible and Necessary Existence Problems in Abstract Argumentation'.

REFERENCES

[1] L. Amgoud and J. Ben-Naim, 'Ranking-based semantics for argumentation frameworks', in *Proceedings of the 7th International Conference on Scalable Uncertainty Management (SUM'13)*, pp. 134–147, (2013).

[2] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G.R. Simari, M. Thimm, and S. Villata, 'Toward artificial argumentation', *AI Magazine*, **38**(3), 25–36, (October 2017).

[3] *Handbook of Formal Argumentation*, eds., P. Baroni, D. Gabbay, M. Giacomin, and L. van der Torre, College Publications, 2018.

[4] B. De Finetti, *Theory of probability: A critical introductory treatment*, volume 6, John Wiley & Sons, 2017.

[5] Jérôme Delobelle, *Ranking-based Semantics for Abstract Argumentation*, Ph.D. dissertation, Artois University, 2017.

[6] P. Dung, 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *Artificial Intelligence*, **77**(2), 321–357, (1995).

[7] M. Goldszmidt and J. Pearl, 'Qualitative probabilities for default reasoning, belief revision, and causal modeling', *Artificial Intelligence*, **84**(1-2), 57–112, (1996).

[8] J. Heyninck, G. Kern-Isberner, and M. Thimm, 'On the correspondence between abstract dialectical frameworks and nonmonotonic conditional logics', in *Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS-33)*, (2020).

[9] G. Kern-Isberner and M. Thimm, 'Towards conditional logic semantics for abstract dialectical frameworks', in *Argumentation-based Proofs of Endearment - Essays in Honor of Guillermo R. Simari on the Occasion of his 70th Birthday*, College Publications, (2018).

[10] P. Matt and F. Toni, 'A game-theoretic measure of argument strength for abstract argumentation', in *Proceedings of the 11th European Conference on Logics in Artificial Intelligence*, pp. 285–297. Springer, (September 2008).

[11] F. Pu, J. Luo, Y. Zhang, and G. Luo, 'Argument ranking with categoriser function', in *Proceedings of the 7th International Conference on Knowledge Science, Engineering and Management*, pp. 290–301, (2014).

[12] K. Skiba, 'A first idea for a ranking-based semantics using system z', in *Online Handbook of Argumentation for AI (OHAAI)*, (2020). To appear.

[13] W. Spohn, 'Ordinal conditional functions: a dynamic theory of epistemic states', in *Causation in Decision, Belief Change, and Statistics*, 105–134, Kluwer, (1988).

[14] Y. Wu, M. Caminada, and M. Podlaszewski, 'A labelling-based justification status of arguments', *Studies in Logic*, **3**(4), 12–29, (2010).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 9–10, Santiago de Compostela, Spain, August 29-30, 2020.

Advanced Management of Fuzzy Semantic Information

Ignacio Huitzil [1]

Abstract. Managing vague and fuzzy semantic information is a challenging topic in the fields of knowledge engineering and Artificial Intelligence (AI) while there has been some work in the field of fuzzy ontologies, there are still many open problems. In this paper, we briefly overview some advanced features of the management of fuzzy ontologies and fuzzy ontology reasoners. Some highlights presented here are new algorithms to learn fuzzy ontologies, novel reasoning algorithms, new methods to manage imprecise knowledge in mobile devices, and the development of real-world applications as a proof of concept of our developments.

1 INTRODUCTION

In recent years, ontologies have become a standard for knowledge representation. An ontology is an explicit and formal specification of the concepts, individuals, and relationships that exist in some area of interest, created by defining axioms that describe the properties of these entities [11]. They have been successfully used in many applications, making knowledge maintenance, addition of semantics to data, information integration, and reuse of components easier. In the W3C's Semantic Web stack, ontologies are represented by a logic-based language such that knowledge expressed in Web Ontology Language OWL. This language provides a strong mechanism for querying and reasoning with data and its current standard version is OWL 2 [3]. By using a reasoner, one can infer facts which are implicitly contained in an ontology.

Ontologies have proved to be useful in many applications, but they are not appropriate to deal with imprecise and vague knowledge, which is very common in many real-world domains. Given that fuzzy logic and fuzzy set theory are appropriate formalisms to handle such imperfect knowledge [12]. Accordingly, fuzzy ontologies generalize classical (*crisp*) ontologies to allow assertions that are not either true or false, but can be partially true. More precisely, concepts denote fuzzy sets, relations denote fuzzy relations, and axioms and facts may hold to some degree of truth. Knowledge representation with fuzzy ontologies can be done with languages such as Fuzzy OWL 2 [1], while reasoning is supported by reasoners such as fuzzyDL [2].

However, there are still many open problems in fuzzy ontologies, such as developing techniques to build them, developing efficient reasoning algorithms, improving the support in mobile devices, and applying them in real-world problems.

Our contributions include: *(i)* creating and evaluating advanced techniques to help building fuzzy ontologies using automatic machine learning algorithms, *(ii)* improving the access from mobile devices, *(iii)* proposing new efficient reasoning algorithms for fuzzy ontologies, and *(iv)* developing real-world applications using fuzzy ontologies as proof of concept.

[1] University of Zaragoza, Spain, ihuitzil@unizar.es

2 FUZZY MANAGEMENT CONTRIBUTIONS

This section describes our main results on different sides of the management of fuzzy ontologies: learning, reasoning, and applications.

Learning fuzzy ontologies. A common problem is how to obtain the concrete definitions of the fuzzy datatypes. We have developed a tool named Datil that implements algorithm to learn fuzzy datatypes for fuzzy ontologies based on the values of the data properties with numerical ranges [10]. A clustering algorithm provides a set of centroids, which are then used as the parameters to build fuzzy membership functions partitioning the domain of a values of the data property. Datil implements several unsupervised clustering algorithms such as *k-means*, *fuzzy c-means*, and *mean shift*, and supports different input formats, including OWL 2. We have applied Datil to many use cases:

- We categorized users according to their life style routine. Starting from the values of 64 data properties (such as the numbers of steps or calories) for 40 volunteers, collected by using wearables, we learnt 4760 fuzzy datatypes (for example, low number of steps) [10]. Then, we defined categories using linguistic terms (for example, a couch potato walks a low number of steps).
- We computed linguistic summaries of the *biometric features* used in gait recognition [8], as they are more informative for a human than the numerical values. Using fuzzy ontologies with data from 91 volunteers, we obtained linguistic descriptions so we can say, e.g., that the individual to be recognized has a high step length.
- Finally, we learnt the fuzzy datatypes used in a beer recommendation system using a *Fuzzy Beer ontology* with 15317 beers [4]. For example, we learnt fuzzy datatypes for the data property ABU (alcohol level). This makes it possible to state, for example, that Estrella Galicia beer is associated to the linguistic label *NeutralAlcohol* with degree 0.69 by comparing its ABU (5.5) and the linguistic label. An evaluation of the quality of the linguistic labels computed by the three clustering algorithms using the opinions of 46 beer aficionados results showed that mean-shift algorithm gave the closest results to human answers (40% of coincidences).

Another approach to compute the definition of fuzzy datatypes for fuzzy ontologies is to build a single fuzzy datatype from multiple fuzzy datatype definitions (i.e., by experts) by using linguistic aggregation operators. We have developed a tool called Fudge that implements several well-known linguistic aggregation strategies such as the convex combination, linguistic OWA, weighted mean, and fuzzy OWA. We have also proposed and implemented two novel linguistic aggregation operators, based on a left recursive form of the convex combination and of the linguistic OWA. The modular design of this tool makes it easy to support more fuzzy operators. Fudge supports missing data and in some cases can use fuzzy quantifiers to define a vector of weights [7].

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 11–12, Santiago de Compostela, Spain, August 29-30, 2020.

Reasoning. We have extended the reasoning engine fuzzyDL in two directions, to solve the instance retrieval and the realization problems, and to be a serializable and incremental reasoner.

- We proposed specific algorithms to solve the instance retrieval and the realization problems with respect to a fuzzy ontology [6]. Our algorithms are based on a reduction of the number of optimization problems to be solved by merging some of them. Our evaluation shows that the novel algorithm to solve the instance retrieval outperforms the previous one, and that in practice it is common to solve a single optimization problem.
- Reasoning on mobile devices is challenging because of their limited resources. We extended fuzzyDL to make it the first serializable and incremental ontology reasoner (to the best of our knowledge) [9]. This makes it possible to reuse previous inferences computed by another device (e.g., a server in the cloud) without starting from scratch. Essentially, we made all classes serializable, and we implemented some prepossessing tasks that avoid repeating computations. In an empirical test (88 ontologies) we shown that fuzzyDL computes smaller serialized files than other serializable reasoners (JFact). We also proved that reasoning time is smaller with respect to the non-incremental version.

Real-world applications. We have worked on gait recognition, recommendation systems, and blockchain.

- Gait recognition implies the automatic human classification from sequences of data about their movement patterns. It has applications in security or medicine. We used fuzzy ontologies to represent sequences of Microsoft Kinect gait data and some biometric features [8]. We built a new dataset, more reusable and interpretable, with 91 ontologies. We also proposed a novel recognition algorithm for straight line walks. It is based on fuzzy logic similarity and a voting scheme to aggregate the values obtained for each step of the sequence. An evaluation showed that out system outperforms existing algorithms. We also faced the problem of the identification of new unknown individuals.
- Beer market is a hot topic which is receiving a notable attention in the last years. We developed GimmeHop, a beer recommendation system for Android mobile devices using fuzzy ontologies and semantic reasoners [4]. GimmeHop is able to deal with user context (in particular, user location) by using fuzzy hedges, with user preferences by using weighted mean aggregation, and with incomplete data by using quantifier-guided OWA to provide weighting vectors with different sizes. We performed an extensive evaluation of several features of the system, including data traffic, running time, quality of the recommendations and quality of the linguistic labels (obtained using Datil). Our experiments show that remote reasoning is feasible and efficient. in terms of data traffic and time. Local reasoning is only feasible if we limit the number of individuals: the tested devices were able to support around 3000 beers.
- We proposed a novel architecture to integrate fuzzy ontologies and blockchain technologies [5]. This way, users can model flexible restrictions (e.g., the price is approximate) using fuzzy sets. The main advantage is that it is possible to develop smart contracts where a partial agreement between the involved parts (e.g., the seller and the buyer) is possible. Computing the partial agreement is reduced to solving a fuzzy ontology reasoning task (the best satisfiability degree of a combination of fuzzy concepts representing the constraints of each of the parts). This approach has four fuzzy ontologies, a schema ontology, a personal ontology to represent each part and a common ontology that contains the agreed values of the smart contract. The architecture is developed on Ethereum, using fuzzyDL reasoner to compute the partial agreements and InterPlanetary File System to store the common ontology.

3 CONCLUSIONS AND FUTURE WORK

We have presented some approaches in the management of fuzzy semantic information, mainly in learning elements for fuzzy ontologies, solving reasoning tasks, and developing real-world applications. Regarding fuzzy ontology learning, we have described Datil and Fudge tools. Regarding reasoning, we have extended fuzzyDL semantic reasoner, and discussed reasoning on mobile devices. We have also discussed applications of fuzzy ontologies in security, recommendation systems, e-commerce, and decision making.

Future work in fuzzy ontology learning includes improving Datil by adding more sophisticated clustering algorithms. A challenge of our applications is to go from a prototype to production stage. We are also interested in applying both Datil and Fudge in some more real-world domains. For fuzzyDL resoner, we would like to support local reasoning on Android devices.

ACKNOWLEDGEMENTS

We were partially supported by DGA/FEDER 2014–2020 and by the project TIN2016-78011-C4-3-R (AEI/ FEDER, UE). All my gratitude to Fernando Bobillo by his mentoring and patient in this work.

REFERENCES

[1] Fernando Bobillo and Umberto Straccia, 'Fuzzy ontology representation using OWL 2', *International Journal of Approximate Reasoning*, **52**(7), 1073–1094, (2011).
[2] Fernando Bobillo and Umberto Straccia, 'The fuzzy ontology reasoner fuzzyDL', *Knowledge-Based Systems*, **95**, 12–34, (2016).
[3] Bernardo Cuenca-Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter F. Patel-Schneider, and Ulrike Sattler, 'OWL 2: The next step for OWL', *Journal of Web Semantics*, **6**(4), 309–322, (2008).
[4] Ignacio Huitzil, Fernando Alegre, and Fernando Bobillo, 'GimmeHop: A recommender system for mobile devices using ontology', *Fuzzy Sets and Systems*, (2020).
[5] Ignacio Huitzil, Álvaro Fuentemilla, and Fernando Bobillo, 'I can get some satisfaction: Fuzzy ontologies for partial agreements in blockchain smart contracts', in *Proceedings of the 29th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2020)*. IEEE Press, (2020).
[6] Ignacio Huitzil, Jorge Bernad, and Fernando Bobillo, 'Algorithms for instance retrieval and realization in fuzzy ontologies', *Mathematics*, **8**(2), 154:1–16, (Jan 2020).
[7] Ignacio Huitzil, Fernando Bobillo, Juan Gómez-Romero, and Umberto Straccia, 'Fudge: Fuzzy ontology building with consensuated fuzzy datatypes', *Fuzzy Sets and Systems*, (2020).
[8] Ignacio Huitzil, Lacramioara Dranca, Jorge Bernad, and Fernando Bobillo, 'Gait recognition using fuzzy ontologies and Kinect sensor data', *International Journal of Approximate Reasoning*, **113**, 354–371, (2019).
[9] Ignacio Huitzil, Umberto Straccia, Carlos Bobed, Eduardo Mena, and Fernando Bobillo, 'The serializable and incremental semantic reasoner fuzzyDL', in *Proceedings of the 29th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2020)*. IEEE Press, (2020).
[10] Ignacio Huitzil, Umberto Straccia, Natalia Díaz-Rodríguez, and Fernando Bobillo, 'Datil: Learning fuzzy ontology datatypes', in *Proceedings of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2018), Part II*, volume 854 of *Communications in Computer and Information Science*, pp. 100–112. Springer, (2018).
[11] Steffen Staab and Rudi Studer, *Handbook on Ontologies*, International Handbooks on Information Systems, Springer, 2004.
[12] Lotfi A. Zadeh, 'Fuzzy sets', *Information and Control*, **8**, 338–353, (1965).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 11–12, Santiago de Compostela, Spain, August 29-30, 2020.

Handling Uncertainty and Risk in Privacy Prediction

Gönül Aycı [1]

Abstract. Online social network users frequently share personal information online. While each post is targeted to a certain audience, it is not always easy to judge what the privacy implications of shared content will be. To ensure that privacy is preserved, each user has to think through these implications before sharing content, which is difficult at best. Recent work advocate use of intelligent systems that can help people preserve their privacy by helping users decide whether the content is private or not so that the user can take an action accordingly; e.g., only share with family as opposed to publicly. Most of these approaches are centralized in that they use the data that has been shared by many users to make recommendations. First of all, we recommend privacy labels for each content. In our proposed text-based approach, we use Natural Language Processing (NLP) techniques such as Term Frequency-Relevance Frequency (TF-RF) while extracting features. At the learning phase, we use Random Forest algorithm. As a classification result of learning models, misclassified images can have different costs. Our second goal is to develop uncertainty and risk models for prediction results of content before sharing. Uncertainty and risk are significant concepts to preserve the privacy of users by deciding privacy settings and identifying the privacy preferences of users. Finally, we would like to explain the decisions of models. Explainability is important for understanding the reasons behind the predictions of proposed algorithms. This part is not clear now. We can explain decisions using such as rules of machine learning algorithms.

1 INTRODUCTION

Privacy is a significant issue for humanity throughout history. Alan Westin defines privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated" [6]. Several studies suggest that before the explosive growth of the Internet, individuals have always had a desire for privacy [10, 14]. However, in recent years, an interest in social media is increasing, preserving the privacy of users become a popular issue. Users share content and interact with each other in OSNs. Nowadays, there are more than 300 web sites collecting the information of more than 400 million registered users [1], and the most popular and the biggest OSNs are Facebook which has over 2.38 billion monthly active users in the worldwide as of March 31, 2019 [2], and Twitter which has 326 million monthly active members as of October 26, 2018 [3].

Privacy of OSNs users is compromised by using social and content sharing platforms [17]. People tend to share text-based content such as tweets or comments and visual-based such as images or videos. One of the most shared content types is an image in OSNs. While sharing images, users can specify different types of relationships "friend", "family," "colleague", etc. Learning the privacy preferences of users from images is a popular and challenging topic because of the subjectivity of privacy. Privacy of images can not be subjective, we can learn privacy patterns and preferences of users using shared images in OSNs. For instance, Zerr et al. [15, 16] conduct a study known as the PicAlert dataset. PicAlert dataset consists of Flickr images that are labeled as private, public, or undecidable by external viewers. An image is considered to be "undecidable" if the user could not decide at first glance (the annotator needs to display metadata of the image or use a link to show original context), or it can be "private" if it belongs to the private sphere (e.g. self-portraits, family, friends, your home) or contains objects that the user can not be shared with the world (e.g. a private email), or otherwise, the rest is considered to be "public". One of the most challenging parts of conduct our study is to find a useful dataset. In our studies, we use the PicAlert dataset. It is seen that 17% of the images have conflicting labels from annotators and only 0.014% of images labeled as undecidable.

Researchers propose visual-based and tag-based approaches for predicting the privacy of images. In visual-based studies, Squicciarini et al. [12, 13] use SIFT, RGB features, and user-annotated tags, and Ashwini [4] explores features and generates tags using a pretrained object CNN for the privacy prediction task. In tag-based studies, user annotated tags are used for access control policies [7], and tool generated tags are used for privacy prediction of images [8]. Kurtan and Yolum [8] propose an agent-based approach that predicts the binary privacy settings of images such as public or private using their content tags. Their approach constructs an internal tag table and an external tag table to learn the privacy preferences of users.

Besides privacy prediction of images, uncertainty, and risk of the learning algorithm's decision are greatly important. All of the wrong decisions are not equal. When the learning algorithm makes wrong decisions, it can cause catastrophic, and costly results. If the learning algorithm classifies private content as public, it can violate her privacy and destroys the quality of her life. We will develop a model of the uncertainty and risk for the privacy prediction task. One of the candidate research questions is that can we calculate a personalized risk? When the image will be misclassified and decide whether to share the content according to the level of the risk and uncertainty values.

2 NOVELTY

In our AI4P ECAI workshop paper, we propose an approach that learns the privacy preferences of the user from tags. Different from the previous privacy prediction models, the proposed approach uses the Term Frequency - Relevance Frequency (TF-RF) method [9] to extract features and classifies the image using a machine-learning algorithm such as the Random Forest (RF) algorithm [5]. When a tag does not valid in the training set, the model suggests a value for the unknown tag based on cosine-similarity metric using vector repre-

[1] Bogazici University, Turkey, email: gonul.ayci@boun.edu.tr

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 13–14, Santiago de Compostela, Spain, August 29-30, 2020.

sentations of the tag. This method reaches acceptable accuracy scores when there are only five tags per image.

We will develop this model using visual features of contents with deep-learning models. There are some studies using one of the popular deep-learning architecture which is CNN [4]. However, we desire to improve by integrating uncertainty and risk into the proposed model. Uncertainty and risk are two significantly important concepts to decide privacy settings, identify the privacy preferences of users, and affect the happiness of users.

3 IMPACT

OSNs user tends to share information about herself. Current social networking sites allow users to change their privacy preferences however OSNs users can have difficulties in managing their privacy settings. There are some HCI studies about this issue in the literature. For instance, some users thought that Facebook's grouping and privacy feature too confusing or difficult to use, and some of them do not trust Facebook [11]. The proposed model will be able to automatically predict privacy settings of the content instead of her while she is sharing content in OSNs. Whenever she desires to use it, the model can decide whether the content will be shared or not.

When the content has tags, we want to develop a tool that will be able to explain her privacy preferences. We are curious about what is private to OSN users? What are the reasons behind the learning algorithm decide whether the content is private or public? In which case the learning algorithm does not decide? Is it possible to develop a human-understandable tool for this purpose?

ACKNOWLEDGEMENTS

I would like to thank to my Ph.D. advisor and co-advisor, Dr. Arzucan Özgür and Dr. Pınar Yolum, for providing guidance and feedback throughout this research.

REFERENCES

[1] List of social networking websites. https://en.wikipedia.org/wiki/List_of_social_networking_websites/.
[2] Top-15 valuable facebook statistics. https://zephoria.com/top-15-valuable-facebook-statistics/.
[3] Twitter statistics. https://www.fastcompany.com/90256723/twitters-q3-earnings-by-the-numbers.
[4] Tonge Ashwini, *Identifying Private Content for Online Image Sharing*, Ph.D. dissertation, 2019.
[5] Leo Breiman, 'Random forests', *Machine learning*, **45**(1), 5–32, (2001).
[6] Mireille Hildebrandt, 'Defining profiling: a new type of knowledge?', in *Profiling the European citizen*, 17–45, Springer, (2008).
[7] Peter Klemperer, Yuan Liang, Michelle Mazurek, Manya Sleeper, Blase Ur, Lujo Bauer, Lorrie Faith Cranor, Nitin Gupta, and Michael Reiter, 'Tag, you can see it!: using tags for access control in photo sharing', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 377–386. ACM, (2012).
[8] Abdurrahman Can Kurtan and Pinar Yolum, 'Pelte: Privacy estimation of images from tags', in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1989–1991. International Foundation for Autonomous Agents and Multiagent Systems, (2018).
[9] Man Lan, Chew Lim Tan, and Hwee-Boon Low, 'Proposing a new term weighting scheme for text categorization', in *AAAI*, volume 6, pp. 763–768, (2006).
[10] Ferdinand David Schoeman, *Philosophical dimensions of privacy: An anthology*, Cambridge University Press, 1984.
[11] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber Lynn McConahy, Jason Wiese, and Lorrie Faith Cranor, 'The post that wasn't: exploring self-censorship on facebook', in *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 793–802. ACM, (2013).
[12] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi, 'Toward automated online photo privacy', *ACM Transactions on the Web (TWEB)*, **11**(1), 2, (2017).
[13] Anna C Squicciarini, Cornelia Caragea, and Rahul Balakavi, 'Analyzing images' privacy for the modern web', in *Proceedings of the 25th ACM conference on Hypertext and social media*, pp. 136–147. ACM, (2014).
[14] Jose M Such, Agustín Espinosa, and Ana García-Fornes, 'A survey of privacy in multi-agent systems', *The Knowledge Engineering Review*, **29**(3), 314–344, (2014).
[15] Sergej Zerr, Stefan Siersdorfer, and Jonathon Hare, 'Picalert!: a system for privacy-aware image classification and retrieval', in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2710–2712. ACM, (2012).
[16] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova, 'Privacy-aware image classification and search', in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 35–44. ACM, (2012).
[17] Elena Zheleva and Lise Getoor, 'To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles', in *Proceedings of the 18th international conference on World wide web*, pp. 531–540. ACM, (2009).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 13–14, Santiago de Compostela, Spain, August 29-30, 2020.

On the Automatic Configuration of Algorithms

Marcelo de Souza*

Abstract. This paper proposes a set of techniques for assisting the automatic configuration of algorithms. First, we present a visual tool to analyze the configuration process. Second, we describe some capping methods to speed up the configuration by early stopping poorly performing executions. Finally, we propose a strategy to automatically identify premature convergence in the configuration.

1 INTRODUCTION

Algorithm configuration is the task of finding one or more parameter settings, also called *configurations*, which optimizes the expected performance of a target algorithm for a particular set of problem instances. A configuration scenario is composed by the parameter space Θ, a set of training instances Π, and a performance metric $c(\theta, \pi)$ of executing the target algorithm under configuration $\theta \in \Theta$ on instance $\pi \in \Pi$. We usually define c as the running time of the execution or the cost of the best found solution, for decision and optimization problems, respectively.

There are different tools that automate the algorithm configuration process. This paper focuses on the irace configurator [1], which implements an iterated racing procedure to explore the parameter space. Given the configuration scenario $\langle \Theta, \Pi, c \rangle$ and a computational budget B, irace iteratively samples a set of configurations and evaluates them using racing. Statistical tests are used to eliminate configurations which perform worse during the racing phase. The surviving (elite) configurations are used by the probabilistic models to sample new configurations for the next iteration. These steps are repeated until the budget B is exhausted. B is a maximum number of evaluations or an execution time limit.

Some difficulties arise when automatically configuring algorithms. First, it is not easy to analyze the configuration process. Mistakes in the design of the configuration scenario (e.g. choosing non-representative training instances or using a too large budget) are hard to detect. Second, configuration is costly, since many candidates must be executed on different instances, and each execution is often time consuming. Finally, the sampling models of irace can converge in an intermediate point of the configuration, then the newly generated candidates are very similar to those already evaluated, loosing diversity. The current convergence detection implemented in irace is not effective in most cases.

We present strategies to deal with the aforementioned problems. First, we propose a visualization tool to analyze the configuration process (Sec. 2). Second, we present several capping methods to avoid wasting time evaluating poor performers (Sec. 3). Finally, we propose a new approach for detecting the premature convergence of the sampling models (Sec. 4).

2 VISUALIZATION TOOL

We propose a tool that produces visual representations of the configuration process performed by irace. Figure 1 presents an example

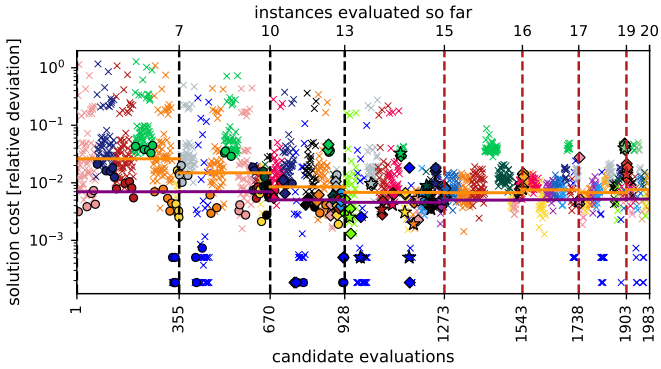

Figure 1. Visualization of the automatic configuration of ACOTSP.

of the main plot for the configuration of ACOTSP, an ant colony optimization algorithm for the traveling salesperson problem. The plot presents all candidate evaluations and indicates the beginning of each iteration (vertical dashed lines) with the corresponding number of candidate evaluations performed so far. Iterations with soft restart are presented in red. Different instances are identified by distinct colors and executions of elite configurations are represented using different markers (○ for elites of the current iteration, ◇ for elites of the last iteration, and ☆ for the best found configuration). For each iteration, we compute the expected performance of both elite and non-elite configuration sets as the median of the results obtained so far. When certain configurations were not evaluated on a given instance, we replace the missing result values by the worst performance of the elite configurations of the corresponding iteration.

The provided visualization gives an overview of the configuration process and allows one to see how the candidates evolve over the iterations. We can also observe how the computational budget is used in each iteration and compare the performance of elite and non-elite configurations. Finally, we can identify possible mistakes in the design of the configuration scenario. For example, Figure 1 shows that the instance represented in green must be hard, since all configurations present almost the same bad performance on it. On the other hand, the instance represented in blue must be easy, since most of the configurations present good performance on it, in comparison to the performances on other instances. Besides that, we observe that from the fourth iteration on, the performance of elite and non-elite configurations does not improve, at the same time that convergence has been detected in all subsequent iterations, indicating a premature convergence of the configuration process.

We provide additional features to allow the user to check more detailed information about each execution by hovering the cursor on the corresponding point. We also allow the user to control the visualization (e.g. by expanding a desired area or moving the plot) and configure the output settings (e.g. by hiding elements or exporting the plot). We also provide a second plot to see the performance of each elite configuration on the test instances, given that the test option was used when running irace. The tool is available at `https://github.com/souzamarcelo/cat`.

* Santa Catarina State University, Federal University of Rio Grande do Sul, Brazil (marcelo.desouza@udesc.br).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 15–16, Santiago de Compostela, Spain, August 29-30, 2020.

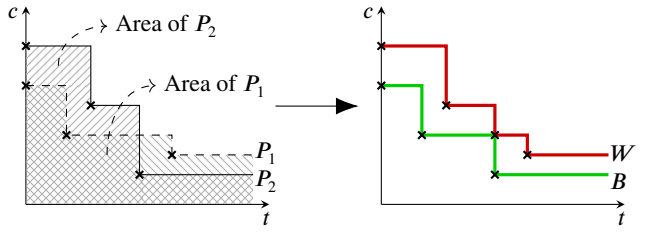

Figure 2. Original (left) and aggregated (right) performance profiles.

3 CAPPING METHODS

During the configuration, a considerable effort is usually spent evaluating poor candidates. For example, observe all the poorly performing executions of the first iterations in Figure 1. We introduce a set of capping methods to speed up the configuration of optimization algorithms. They use previously seen executions to determine a performance envelope, i.e. a minimum required performance, and then use it to evaluate how good new executions are. When the conditions of the envelope are violated, the execution is stopped.

The capping methods behave as follows. The performance envelope is computed before evaluating a new configuration on a given instance. Each previous execution is represented by its performance profile, i.e. the solution cost as a function of the running time, where $P(t) = c$ gives the cost c of the best found solution after running the algorithm for a time t (or any other measure of computational effort). Figure 2 presents two performance profiles (P_1 and P_2) to be aggregated into the performance envelope. We propose two types of envelope. The *profile-based envelope* is given by a performance profile and can be obtained by applying the worst (W) or best (B) aggregation functions to the set of performance profiles. The worst aggregation function selects the pointwise maximum solution cost among the performance profiles for each value of time t, then $W(P_1, \dots, P_n)(t) = \max\{P_1(t), \dots, P_n(t)\}$. Analogously, the best aggregation function selects the pointwise minimum solution cost, then $B(P_1, \dots, P_n)(t) = \min\{P_1(t), \dots, P_n(t)\}$. Figure 2 presents the aggregated performance profiles for both functions on the right side.

The *area-based envelope* considers the area defined by the performance profiles, instead of dealing directly with them. The area of a performance profile P is given by $A_P = \int_0^{t_f} P(t)\,\mathrm{d}t$, where t_f is the cut-off time for finished executions or the current effort of an execution in progress. In this approach, the performance envelope is given by a maximum area available for the next execution. Similarly to the previous strategies, the maximum area is obtained by the worst aggregation function $W(P_1, \dots, P_n) = \max\{A_{P_1}, \dots, A_{P_n}\}$, or the best aggregation function $B(P_1, \dots, P_n) = \min\{A_{P_1}, \dots, A_{P_n}\}$.

We applied the proposed capping methods for the configuration of the following algorithms: ACOTSP, HEACOL (a hybrid evolutionary algorithm for the graph coloring), TSBPP (a tabu search for the bin packing problem), and HHBQP (a hybrid heuristic for the unconstrained binary quadratic programming). We executed irace 20 times for each capping method and computed the mean effort savings in comparison to configuring without capping. We also executed the resulting algorithms 5 times to compute the mean relative deviation from the best known solutions (for HHBQP, we computed the mean absolute deviation). Table 1 presents the results when using no capping (first line), for profile- and area-based approaches (P and A) and both aggregation functions (W and B). The best values of each scenario are in boldface. We observe that all capping methods produce significant effort savings, while the quality of the final configurations does not degrade. As expected, the best aggregation function is more aggressive, producing greater effort savings but worse final configurations in comparison to the worst aggregation function. More detailed results and the source code of these and additional capping strategies can be found at `https://capopt.github.io`.

Table 1. Mean effort savings and mean deviation for each capping method.

Cap.	ACOTSP		HEACOL		TSBPP		HHBQP	
	sav.	r. dev.	sav.	r. dev.	sav.	r. dev.	sav.	a. dev.
-	-	**0.33**	-	**4.14**	-	1.31	-	49.72
PW	59.7	0.37	61.3	4.22	22.6	**1.25**	44.3	65.16
PB	**77.7**	0.52	**74.8**	4.48	38.1	1.27	**74.9**	58.38
AW	26.8	0.35	27.2	4.18	12.4	1.28	17.3	**46.97**
AB	52.7	0.38	47.0	4.18	**41.4**	1.35	65.9	68.56

4 CONVERGENCE DETECTION

The current method of convergence detection implemented in irace is based on a distance measure between elites and each generated configuration. If the distance is less than a threshold, the associated sampling models are restarted. We present an alternative approach to detect convergence, which compares the performance of the configurations of the current and previous iterations, instead of analyzing the configurations individually. If the difference in the observed quality (i.e. the mean running time or the mean relative deviation from a lower bound) of both iterations is smaller than a threshold, the sampling models are restarted.

We analyzed previous runs of irace and applied our approach to the available data. We observe that the proposed method is able to detect the convergence almost whenever the current method detects. Besides that, the proposed approach is able to detect convergence in some cases when no improvement is reached, but the current method does not detect any convergence. We plan to extend this method to consider not only the observed performances, but the evolution of the probability distributions used to sample new configurations.

5 CONCLUDING REMARKS

The preliminary experiments indicate that the proposed methods are effective for improving the automatic configuration of algorithms. The visualization provides useful information and can be helpful to understand the configuration process and to identify possible mistakes in the configuration scenario. The capping methods led to a significant reduction in the configuration time for different scenarios. This allows us to scale the automatic methods to challenging configuration tasks. Finally, an effective method to detect convergence can help to diversify the parameter space exploration when appropriate, and then make a better use of the available computational budget.

For the next steps, we plan to extend the experiments. We will use the visualizations to analyze an extensive set of configuration scenarios, identifying common made mistakes and how to visualize them. We also want to apply the capping methods using the total configuration time as budget and allowing irace to use the saved effort to further explore the parameter space.

REFERENCES

[1] Manuel López-Ibáñez, Jérémie Dubois-Lacoste, Leslie Pérez Cáceres, Thomas Stützle, and Mauro Birattari, 'The irace package: Iterated racing for automatic algorithm configuration', *Operations Research Perspectives*, **3**, 43–58, (2016).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 15–16, Santiago de Compostela, Spain, August 29-30, 2020.

Evolving priority rules for scheduling problems by means of hyperheuristics

Francisco J. Gil-Gala[1]

Abstract. On-line scheduling is often required in a number of real-life settings. This is the case of distributing charging times for a large fleet of electric vehicles arriving stochastically to a charging station in which there are a number of physical and power constraints. In this thesis, we consider a scheduling problem derived from a situation of this type: the one machine scheduling problem with variable capacity over time and tardiness minimization, denoted $(1, Cap(t)||\sum T_i)$. The goal is to develop new priority rules to improve the results from some classical ones as the Apparent Tardiness Cost (ATC) rule. To this end, we propose to use Genetic Programming (GP) and state space search. We have already developed some prototypes showing high capacity to obtain high performing rules.

1 INTRODUCTION

Scheduling problems appear profusely in many fields of industry and management and stand out for their high computational complexity (generally they are NP-hard). For this reason, they often require the use of advanced algorithms and techniques from Artificial Intelligence. This thesis focuses on a problem of this class, denoted $(1, Cap(t)||\sum T_i)$, in which a number of jobs must be scheduled on a single machine whose capacity varies over time, with the goal of minimizing the *total tardiness* objective function. Figure 1 shows an example.

This problem was introduced in [7] in the context of scheduling the charging times of a large fleet of Electric Vehicles (EVs). Specifically, solving the Electric Vehicle Charging Scheduling Problem (EVCSP) tackled in [7] requires solving a number of instances of the $(1, Cap(t)||\sum T_i)$ problem over time. Due to the computational intractability of this problem and the tight real-time requirements of the EVCSP, *on-line scheduling* represents the most (if not the only) suitable approach to the $(1, Cap(t)||\sum T_i)$ problem. In [7], it is solved by means of the *Apparent Tardiness Cost* (ATC) priority rule, commonly used in the context of scheduling with tardiness objectives.

The terms *Dispatching Rule* (DR) and *Priority Rule* (PR) are commonly used in the scheduling literature to refer to "a simple heuristic that derives a priority index of a job from its attributes" [1]. Due to their low computational cost, PRs are well suited for on-line scheduling: the job with the highest priority among those available at a given time is scheduled next.

Priority rules can be defined manually by experts on the problem domain, as it is the case of the ATC rule [8], although it is clear that automatic methods could capture some characteristics of the scheduling problem that are not clear to human experts. For this purpose, *hyper-heuristics* are a suitable choice as they are "heuristics to choose heuristics" [2]. As priority rules may be naturally represented by expressin trees, Genetic Programming (GP) [9] is a common approach. Several authors have adopted the GP paradigm for learning rules for scheduling problems, such as job shop scheduling [11, 6], among others.

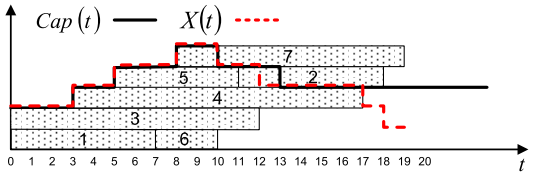

Figure 1. Feasible schedule for an instance of the $(1, Cap(t)||\sum T_i)$ problem with 7 jobs and a machine with capacity varying between 2 and 5 over time.

Figure 2 shows an expression tree representing the classic ATC rule. As usual, it includes a set of function symbols as +,-,*,/,max or exp; a set of problem attributes as d_i (due date of job i), p_i (duration of job i), $\bar{p}$ (average duration of the unscheduled jobs) or $\gamma(\alpha)$ (earliest starting time for the next job); and a set of constant symbols as 1.0 or g (a parameter that must be established). In our study, we consider a set of symbols similar to the ones above.

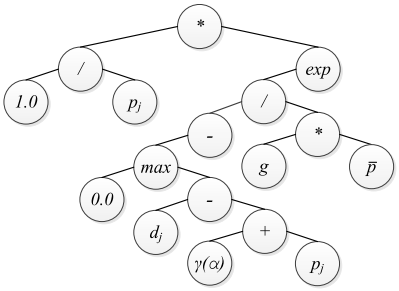

Figure 2. Expression tree representing the ATC rule.

In the next section we describe the thesis proposal. Then, in section 3 we review some of the research work done so far and show some results from the prototypes developed. Finally, in section 4 we summarize the main conclusions that may be drawn from the research work done so far and some ideas for the future.

[1] University of Oviedo, Spain, email: giljavier@uniovi.es

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 17–18, Santiago de Compostela, Spain, August 29-30, 2020.

2 THESIS PROPOSAL

The thesis is carried out within the project TIN2016-79190-R of the research team iScOp;[2] in particular within the general goal aimed at the resolution of scheduling problems with metaheuristics and the more specific one focused on new methods to solve the Electric Vehicle Charging Scheduling Problem (EVCSP) formulted in [7].

The main objective of the thesis is to develop efficient methods to solve the $(1, Cap(t)||\sum T_i)$ problem online and the integration of these methods in the EVCSP solver. More specifically:

1. Analysis of the $(1, Cap(t)||\sum T_i)$ problem and development of schedule generation schemes.
2. Exploiting the use of hyperheuristics, in particular Genetic Programming, to evolve new priority rules to guide the schedule generation schemes.
3. Exploring the use of state space seach paradigm as an alternative to GP in the context of small search spaces.
4. Design of local search procedures to be combined with the above methods.
5. Incorporation of the evolved rules in the EVCSP solver.

3 WORK DONE AND IN PROGRESS

In [3] a GP was proposed to evolve priority rules for the $(1, Cap(t)||\sum T_i)$ problem. The GP follows the strategy proposed in [9], but incorporates new elements adapted to the problem as a smart representation of tree expressions, which allows for limiting the search space, or the restriction to *dimensionally compliant* rules, which are more rational and explainable. Then, a state space search alternative termed *Systematic Search and Heuristic Evaluation* (SSHE) is proposed in [3], which is effective in small search spaces thanks to the use of symmetry breaking rules and a heuristic procedure to evaluate candidate rules. We have also considered the exploitation of ensemble of rules in [5]. Our approach consists in running a small number of rules in parallel and take the result from the best rule as the solution of the ensemble. This method is effective as long as the online requirements are fulfilled. The problem of computing ensembles was formulated as a varaint of the maximum coverage problem and solved by a Genetic Algorithm (GA) from a large pool of rules obtained by other methods.

In the above algorithms, the rules and the ensembles are evaluated on a set of training instances of the $(1, Cap(t)||\sum T_i)$ problem. Then the evolved rules are tested on a set of unseen instances with similar features. Table 1 summarizes some results from the above methods and provides a comparison with some classic hand desiged rules and a Memetic Algorithm (MA) proposed in [10] to solve the problem offline.

Table 1. Summary of results on the benchmark set proposed in [10] produced by the methods proposed in [3, 4, 5].

	Avg. Tardiness	
Rule	Training (50 inst.)	Testing (950 inst.)
SPT	5473.18	5319.81
EDD	1256.84	1291.66
ATC(0.5)	1000.52	1011.33
Best rule (GP)	986.32	992.14
Best rule (SSHE)	993.90	997.05
Best ensemble (GA)	971.32	978.26
MA (30 runs) Best/Avg.	950.18/950.63	959.06/959.51

[2] https://www.di.uniovi.es/iscop

4 CONCLUSION AND FUTURE WORK

We have seen that GP and heuristic search are suitable methods to obtain prioirty rules for the $(1, Cap(t)||\sum T_i)$ problem, and that these rules outperform classical ones as SPT, EDD or ATC in this problem. We have also analyzed the use of ensembles of rules and demonstrated that they may be suitable, but at the cost of increasing the processing time. From the differences between online and offline algorithms, we have seen that there is still room to improve. In this regard, one of the biggest challenges is to devise some neighbourhood structures that can be exploited by a local search algorithm (LSA), which could then be combined with GP or to be exploited to design GRASP like algorithms. The experimental results also encourage further research, as for example devising new optimization algorithms or applying the proposed approach to other scheduling problems.

ACKNOWLEDGEMENTS

This research has been supported by the Spanish Government under research project TIN2016-79190-R and scholarship FPI17 / BES-2017-08203; and by the Principality of Asturias under grant IDI/2018/000176. I would like to thank my PhD supervisor Ramiro Varela and my colleagues María R. Sierra and Carlos Mencía for their valuable help. We are the authors of reference [4], which is included in the main ECAI2020 conference.

REFERENCES

[1] Jürgen Branke, Torsten Hildebrandt, and Bernd Scholz-Reiter, 'Hyper-heuristic evolution of dispatching rules: A comparison of rule representations', *Evolutionary Computation*, **23**(2), 249–277, (2015).
[2] Edmund K. Burke, Matthew R. Hyde, Graham Kendall, Gabriela Ochoa, Ender Özcan, and John R. Woodward, *A Classification of Hyper-Heuristic Approaches: Revisited*, 453–477, Springer International Publishing, Cham, 2019.
[3] Francisco J. Gil-Gala, Carlos Mencía, María R. Sierra, , and Ramiro Varela, 'Evolving priority rules for on-line scheduling of jobs on a single machine with variable capacity over time', *Applied Soft Computing*, **85**, (2019).
[4] Francisco J. Gil-Gala, Carlos Mencía, María R. Sierra, , and Ramiro Varela, 'Exhaustive search of priority rules for on-line scheduling', in *Proceedings of the 2020 Conference on ECAI 2020: 24th European Conference on Artificial Intelligence*, (2020).
[5] Francisco J. Gil-Gala and Ramiro Varela, 'Genetic algorithm to evolve ensembles of rules for on-line scheduling on single machine with variable capacity', in *From Bioinspired Systems and Biomedical Applications to Machine Learning*, eds., José Manuel Ferrández Vicente, José Ramón Álvarez-Sánchez, Félix de la Paz López, Javier Toledo Moreo, and Hojjat Adeli, pp. 223–233, Cham, (2019). Springer International Publishing.
[6] Emma Hart and Kevin Sim, 'A hyper-heuristic ensemble method for static job-shop scheduling', *Evolutionary Computation*, **24**(4), 609–635, (2016).
[7] Alejandro Hernández-Arauzo, Jorge Puente, Ramiro Varela, and Javier Sedano, 'Electric vehicle charging under power and balance constraints as dynamic scheduling', *Computers & Industrial Engineering*, **85**, 306 – 315, (2015).
[8] C. Koulamas, 'The total tardiness problem: Review and extensions', *Operations Research*, **42**, 1025–1041, (1994).
[9] John R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, 1992.
[10] Carlos Mencía, María R. Sierra, Raúl Mencía, and Ramiro Varela, 'Evolutionary one-machine scheduling in the context of electric vehicles charging', *Integrated Computer-Aided Engineering*, **26**(1), 1–15, (2019).
[11] Su Nguyen, Yi Mei, Bing Xue, and Mengjie Zhang, 'A hybrid genetic programming algorithm for automated design of dispatching rules', *Evolutionary Computation*, **27**(3), 467–496, (2019).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 17–18, Santiago de Compostela, Spain, August 29-30, 2020.

A Metareasoning Framework for Planning and Execution in Autonomous Systems

Justin Svegliato[1]

Abstract. *Metareasoning*, a particularly effective computational approach to bounded rationality, is the process by which an autonomous system optimizes its own planning and execution in order to act effectively in its environment. The need for metareasoning has become critical to autonomous systems due to the uncertainty about the range of their potential circumstances and the limitations of their reasoning capabilities. My thesis develops a novel metareasoning framework for monitoring and controlling both the *planning* processes and *execution* processes of autonomous systems. The planning module employs efficient metareasoning techniques that relax unrealistic assumptions of earlier work while the execution module employs robust metareasoning techniques that resolve a range of exceptions and maintain a level of safety. The result is a modular and nonmyopic metareasoning framework that monitors and controls the planning and execution of autonomous systems.

1 INTRODUCTION

It has long been recognized that autonomous systems cannot be capable of perfect rationality due to the intractability of optimal decision making in complex domains [5]. As a result, there have been substantial efforts to develop computational approaches to bounded rationality [4]. *Metareasoning*, a particularly effective computational paradigm for bounded rationality, enables an autonomous system to optimize its own planning and execution processes in order to act effectively in its environment. This enables the autonomous system to handle any uncertainty about the range of its potential circumstances and the limitations of its reasoning capabilities. Consequently, due to the growth in the complexity of autonomous systems in recent years, metareasoning has become critical to automated decision making.

There has been considerable progress in developing metareasoning techniques for monitoring and controlling the planning processes of autonomous systems. For example, a recent method identifies the best algorithm to solve a problem among a portfolio of algorithms by compiling a model with a limited number of features to predict the efficiency and accuracy of the each algorithm [3]. Another recent method selects the next computation, specifically the next simulation, to be performed by Monte Carlo search techniques by representing the decision as a Bayesian selection problem that maximizes the value of information [2]. There are also many methods that determine when to interrupt an anytime algorithm and act on the current solution by using a profile that represents the performance of the anytime algorithm [1]. However, despite these advances, developing metareasoning techniques for monitoring and controlling the execution processes of autonomous systems has not seen much attention.

[1] College of Information and Computer Sciences, University of Massachusetts Amherst, USA, email: jsvegliato@cs.umass.edu

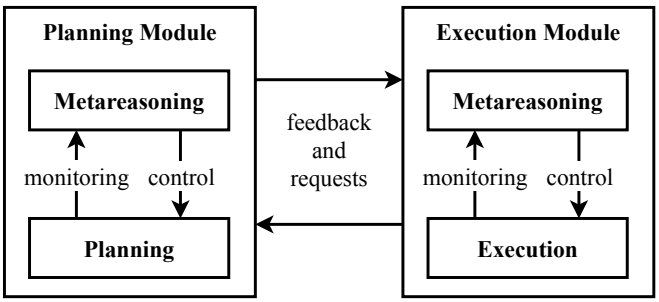

Figure 1. A metareasoning framework for monitoring and controlling the planning processes and execution processes of an autonomous system.

In my thesis, I develop a metareasoning framework for autonomous systems that improves the efficiency of meta-level control for planning processes and expands the scope of meta-level control to execution processes. The framework is composed of a pair of modules shown in Figure 1. The *planning module* enables the system to monitor and control its planning processes to generate policies for acting effectively in its environment within an acceptable amount of time. My thesis proposes methods used by the planning module that relaxes unrealistic assumptions made by earlier work. The *execution module* enables the system to monitor and control its execution processes to act appropriately on the policies generated by the planning module. My thesis introduces methods used by the execution module that resolve a range of exceptions and maintain a level of safety.

The general aim of my thesis is therefore to build a metareasoning framework, which is composed of a planning module and an execution module, that makes the following contributions.

1. **Online Performance Prediction.** Determine the optimal stopping point of an anytime algorithm by predicting the performance of the algorithm online to avoid relying on substantial offline preprocessing needed to compile and maintain a performance profile.
2. **Model-Free Meta-Level Control.** Employ reinforcement learning methods to learn and adapt a policy that indicates when to interrupt an anytime algorithm and act on the current solution to handle changes in the parameters of meta-level control.
3. **Adjustable Anytime Algorithms.** Adjust the hyperparameters of an anytime algorithm at runtime to attain the best solution in the shortest amount of time using deep reinforcement learning.
4. **Exception Recovery.** Resolve a range of exceptions that prevent an autonomous system from completing its task by recovering from unanticipated scenarios during execution.
5. **Safe Operation.** Maintain a level of safety as an autonomous system completes a task by periodically monitoring and proactively avoiding potentially unsafe situations during execution.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 19–20, Santiago de Compostela, Spain, August 29-30, 2020.

2 COMPLETED WORK

We describe the contributions that have been made toward the planning and execution modules of the metareasoning framework below.

Contributions 1 and 2. Autonomous systems rarely have enough time to determine the optimal solution to real world decision-making problems. To generate an acceptable solution under strict time constraints, an autonomous system often uses an anytime algorithm that gradually improves the quality of a solution as it runs and returns the current solution if it is interrupted. However, to exploit the trade-off between solution quality and computation time, it must decide when to interrupt the anytime algorithm and act on the current solution.

Existing metareasoning techniques that monitor and control anytime algorithms rely on planning with a model, called a *performance profile*, that describes the performance of a given anytime algorithm solving a specific problem on a particular system [1]. This model is compiled offline before the activation of meta-level control by using the anytime algorithm to solve thousands of instances of the problem on the system. Planning with a model, however, imposes many assumptions often violated by autonomous systems in the real world. First, there must be enough time for offline compilation of the performance profile of the algorithm. Second, the settings of the algorithm across every problem instance must be the same. Third, the distribution of problem instances solved by the algorithm must be known and fixed. Fourth, the CPU and memory conditions of the system executing the algorithm must be static.

Addressing these unrealistic assumptions, we propose two metareasoning approaches in recent work [9, 10, 7]. Both approaches monitor the performance of the anytime algorithm and estimate the stopping point at runtime by expressing the state of computation in terms of solution quality and computation time. However, the first approach predicts the performance online *with* a model by using a performance predictor while the second approach learns the performance through experience *without* a model by using reinforcement learning. For each approach, we empirically show their effectiveness on a set of common benchmark domains and a mobile robot domain.

Contribution 4. Due to the complexity of the real world, autonomous systems use decision-making models that rely on simplifying assumptions to make them computationally tractable and feasible to design. However, since these limited representations cannot fully capture the domain of operation, an autonomous system may encounter unanticipated scenarios that cannot be resolved effectively.

A simple approach to ensuring the necessary conditions of normal operation is to place the entire responsibility on the operator deploying the autonomous system. However, while relying on human judgment can improve performance, it is desirable to limit human involvement when the conditions of normal operation are violated. Recent work in automated exception recovery has focused on fault diagnosis—detecting and identifying faults—during normal operation. For instance, many approaches diagnose faults by using particle filters or multiple model estimation with neural networks [11]. While these approaches *detect* and *identify* exceptions, they do not offer a way to *handle* exceptions without human assistance. Building on recent work in fault diagnosis, our goal is to develop an exception recovery framework that detects, identifies, and handles exceptions.

We offer an approach to *introspective autonomous systems* in recent work [8]. This system uses belief space metareasoning to recover from exceptions by interleaving a main decision process with exception handlers based on a belief over exceptions that can arise during normal operation. We show that an introspective autonomous vehicle is effective in simulation and on a fully operational prototype.

3 FUTURE WORK

This year, we will make the remaining contributions toward the planning and execution modules of the metareasoning framework below.

Contribution 3. Although our contributions have focused on how to determine when to interrupt an anytime algorithm and act on the current solution, adjusting its internal parameters to optimize performance has not been explored yet. We are developing a metareasoning approach that learns how to adjust its internal parameters by using deep reinforcement learning with a rich state of computation. The state of computation could include features specific to the problem, algorithm, or system. Formally, this approach learns a policy of an MDP with states representing the state of computation and actions representing the internal parameters of the anytime algorithm. We will evaluate our approach by learning how to adjust the weight of Anytime A* and the growth factor of RRT*.

Contribution 5. Ensuring safety is critical to autonomous systems that operate in the real world. We are developing an approach that monitors and controls the execution of an autonomous system to maintain and restore a degree of safety: a set of meta-level monitors address any potential safety problems as a main decision process completes a task. Formally, the main decision process is an MDP that recommends actions that complete a task while each meta-level monitor is an MDP that recommends actions that constrain the actions of the main decision process. The objective of each meta-level monitor is to maximize the probability of remaining in a safe region of the state space while minimizing any interference to the main decision process. We will evaluate our approach in a simulation with an autonomous vehicle that must navigate a route while encountering intermittent safety issues, such as traction loss or overheating.

ACKNOWLEDGEMENTS

This work was published with a conference paper [6] and supported by NSF grants DGE-1451512, IIS-1724101, and IIS-1813490.

REFERENCES

[1] Eric A. Hansen and Shlomo Zilberstein, 'Monitoring and control of anytime algorithms: A dynamic programming approach', *JAIR*, (2001).
[2] Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony, 'Selecting computations: Theory and applications', *arXiv preprint arXiv:1408.2048*, (2014).
[3] Falk Lieder, Dillon Plunkett, Jessica B. Hamrick, Stuart J. Russell, Nicholas J. Hay, and Thomas L. Griffiths, 'Algorithm selection by rational metareasoning as a model of human strategy', in *NIPS*, (2014).
[4] Stuart J. Russell and Eric H. Wefald, *Do the Right thing: Studies in Limited Rationality*, MIT Press, Cambridge, MA, 1991.
[5] Herbert A. Simon, *Models of Bounded Rationality*, MIT Press, Cambridge, MA, 1982.
[6] Justin Svegliato, Samer Nashed, and Shlomo Zilberstein, 'An integrated approach to moral autonomous systems', in *ECAI*, Santiago de Compostela, Spain, (2020).
[7] Justin Svegliato, Prakhar Sharma, and Shlomo Zilberstein, 'A model-free approach to meta-level control of anytime algorithms', in *ICRA*, Paris, France, (2020).
[8] Justin Svegliato, Kyle Hollins Wray, Stefan J Witwicki, Joydeep Biswas, and Shlomo Zilberstein, 'Belief space metareasoning for exception recovery', in *IROS*, Macau, China, (2019).
[9] Justin Svegliato, Kyle Hollins Wray, and Shlomo Zilberstein, 'Meta-level control of anytime algorithms with online performance prediction', in *IJCAI*, Stockholm, Sweden, (2018).
[10] Justin Svegliato and Shlomo Zilberstein, 'Adaptive metareasoning for bounded rational agents', in *IJCAI Workshop on AEGAP*, Stockholm, Sweden, (2018).
[11] Vandi Verma, Geoff Gordon, Reid Simmons, and Sebastian Thrun, 'Particle filters for fault diagnosis', *RA Magazine*, (2004).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 19–20, Santiago de Compostela, Spain, August 29-30, 2020.

Smart Object Segmentation to Enhance the Creation of Interactive Environments

Marcel Tiator[1]

Abstract. The objective of our research is to enhance the creation of interactive environments such as in VR applications. An interactive environment can be produced from a point cloud that is acquired by a 3D scanning process of a certain scenery. The segmentation is needed to extract objects in that point cloud to, e.g., apply certain physical properties to them in a further step. It takes a lot of effort to do this manually as single objects have to be extracted and post-processed. Thus, our research aim is the real-world, cross-domain, automatic, semantic segmentation without the estimation of specific object classes.

1 INTRODUCTION

The real world can be captured in 3D with methods such as photogrammetry and laser scanning. The acquired data of this methods is used to generate a point cloud from which a 3D mesh can be reconstructed. The resulting mesh can be used for virtual, augmented or mixed reality applications. For instance, we captured different rooms of a hospital to produce a VR tour for the patients. All objects of a room will be connected when transforming a room to a 3D mesh without any processing such that no special interactions with objects are possible. Hence, the objects have to be separated to realize special interactions such as grasping objects in VR. In this work, we consider the extraction of objects from the point cloud domain as, e.g., the level of detail of single objects can be controlled individually during the mesh reconstruction.

The extraction of objects from a point cloud can be done by a segmentation algorithm. On the one hand, the segmentation can be conducted by a classifier where an object with a semantic meaning should be recognized such as in [5]. On the other hand, the segmentation can be applied by a geometric method such as the region growing algorithm or the random sample consensus algorithm [2]. The former classification methods may only are able to extract learned objects accurately. The latter geometric methods are mostly unbiased but often hard to tune as certain expert parameters have to be set. Hence, we apply deep reinforcement learning (DRL) to segment a point cloud to make use of the geometric segmentation methods in combination with neural network agents. The approach for the segmentation of a point cloud is schematically depicted in Figure 1. An agent perceives the point cloud with the progress of the segmentation as observation and chooses an appropriate action. Subsequently, the agent perceives the next observation and a reward that represents the goodness of the action in context of the segmentation. The aim of the approach is that the reward is maximised such that the agent segments a point cloud optimally. By our knowledge, the novelty of our approach is that the agent conducts the segmentation in a sequential environment without the explicit knowledge of classes or expert parameters.

[1] University of Applied Sciences Düsseldorf, Germany, email: marcel.tiator@hs-duesseldorf.de

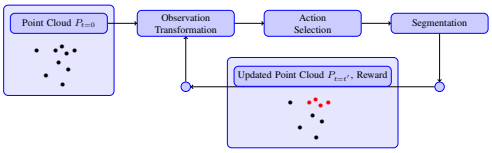

Figure 1. Schematic of the deep reinforcement learning segmentation approach.

2 POINT CLOUD SEGMENTATION

We consider point clouds of the form $P \in \mathbb{R}^{|P| \times D}$ as a matrix of points. There are $|P|$ points in the point cloud and an i-th point $p_i \in P$ has $D \geq 3$ features whereas the first 3 features are spatial features. A segment of a point p_i is characterised by the value $s_i \in \mathbb{N} \cup \{0\}$. The objective of the segmentation of a point cloud is to assign each point p_i to a segment $s_i \neq 0$ such that each segment represents an object. Initially, each point p_i is unsegmented what we encoded with $s_i = 0$. The output of the segmentation is a vector $\hat{s} \in (\mathbb{N} \cup \{0\})^{|P|}$ and we will denote the true segment vector with s. Optimally, the generated segment vector $\hat{s}$ that can be produced by an agent is equal to the true segment vector s. A certain true segment should be denoted as k and an assigned segment that is generated during the segmentation will be denoted as j.

3 RELATED WORK

Liu et al. [3] leveraged DRL for the semantic segmentation of point clouds. An eye window that looks to a point cloud is controlled by a deep q-network. A reward vector is computed by a 3D convolutional neural network (CNN) which also receives the input of the eye window. The output of the 3D CNN is the probability that the eye window contains a target class object. Zhong et al. [6] also used reinforcement learning (RL) in combination with point clouds to solve a semantic segmentation problem. We also use RL in combination with point cloud processing. However, we do not consider a semantic segmentation problem. Instead, the agent should segment the objects without taking care about the explicit class of the object.

4 APPROACHES

Currently, we consider three major components to approach the problem of Section 2:

1. Data: Multiple point clouds with the corresponding true segment vectors for the training of the agent.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 21–22, Santiago de Compostela, Spain, August 29-30, 2020.

2. Mapping: A structure, e.g. a dictionary, that maps an assigned segment vector j to a true segment k. The mapping is used during the segmentation to reward the agent.
3. Environment & Agent: The environment accepts certain actions from an agent, e.g. such as a deep neural network (DNN), to conduct a segmentation step.

4.1 Region growing

We developed an environment that uses the region growing algorithm as first approach. A DNN agent sees a point cloud as observation, estimates the parameters of a region growing algorithm and earns a reward value. The data is generated by sampling points from virtual mesh scenes. Hence, a point to object relationship exists for the training of the agent.

The reward function rewards the agent for correctly assigned points at the end of an episode. The number of unsegmented points u and the number of erroneous assigned points e are counted. To count e, an assigned segment j has to be chosen that represent the true object with segment value k. This is done by taking the mode j^{max} over the distribution of assigned segment values within a true segment k. If a mode j^{max} already exists in the mapping, the next best mode will be selected. If every assigned segment is already mapped, all assigned points within the true segment will be considered as erroneous. In contrast, the mapped assigned segment j within a true segment k will be considered as correct. Hence, all points can be categorised as erroneous, unsegmented or correct such that a reward $r = 1 - \frac{e+u}{|P|}$ can be calculated.

4.2 Perceptual grouping of superpoints

The approach in Section 4.1 has the following disadvantages:

- The point normals have to be calculated for the region growing algorithm which can be error-prone.
- No real world point clouds are used.
- The reward is sparse as it is calculated at the end of the episode.
- If using DNNs such as PointNet [5], a certain number of points have to be sampled from the cloud.

Therefore, we developed an environment that uses point cloud scenes that are generated from the ScanNet data set [1] as the data set contains reconstructed labelled mesh scenes. A DNN observes the point cloud from multiple perspectives as rendered images and segments it in a perceptual grouping task (PGT). Concretely, a point cloud is divided into groups of points, named superpoints, by the voxel cloud connectivity segmentation algorithm [4] that also outputs an adjacency graph of the superpoints. During the PGT, a main and a neighbour superpoint are selected by a certain selection strategy whereas the agent can decide to group them or not. A schematic of the PGT with the superpoint growing selection strategy is depicted in Figure 2. The black rectangles visualise the true segments and the whole green box a point cloud. The coloured isles visualise the superpoints. The grouping suggestions are highlighted with coloured dots in the middle of the coloured isles. The main superpoint has a red dot and the neighbour superpoint a blue dot. The neighbourhood adjacency graph is visualised by the black line connections between the superpoints. After a grouping is confirmed, the neighbours of the neighbour superpoint will also be considered as neighbours of the main superpoint and put on top of a FIFO queue. The main superpoint region is grown till all neighbours are visited. After that, the next main superpoint is chosen and the procedure begins from the beginning. Additionally, if a superpoint region is finished, the reward will be computed if determinable.

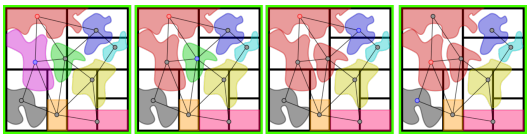

Figure 2. Schematic of the superpoint growing environment to realise the point cloud segmentation as perceptual grouping task.

5 FUTURE WORK

Experiments with an multiview convolutional neural network agent in the environment of Section 4.2 are promising but lead to local optima. This local optima can emerge by assigning the correct number of objects while disregarding the object borders. Moreover, the reward sparsity can be diminished by using heuristics during the grouping process.

To use point clouds from different domains, we also focus on the virtual generation of point cloud scenes such as described in Section 4.1. Currently, we try to optimise the point cloud generation by a ray tracer such that lighting information of the points are included.

ACKNOWLEDGEMENT

Many thanks to my PhD advisors Paul Grimm and Christian Geiger. This research has been funded by the Federal Ministry of Education and Research (BMBF) of Germany in the framework of interactive body-centered production technology 4.0 (German: Interaktive körpernahe Produktionstechnik 4.0 - iKPT4.0) (project number 13FH022IX6).

Marcel Tiator, Christian Geiger, and Paul Grimm, Point Cloud Segmentation with Deep Reinforcement Learning, *in Proceedings of the 24th European Conference on Artificial Intelligence - ECAI 20*, Santiago de Compostela, Spain, (2020). IOS Press.

REFERENCES

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niebner, 'ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes', in *Proceedings of the Conference on Computer Vision and Pattern Recognition - CVPR '17*, Honolulu, Hawaii, (2017). IEEE.

[2] E. Grilli, F. Menna, and F. Remondino, 'A Review of Point Cloud Segmentation and Classification Algorithms', *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS*, **XLII-2/W3**(2W3), 339–344, (feb 2017).

[3] Fangyu Liu, Shuaipeng Li, Liqiang Zhang, Chenghu Zhou, Rongtian Ye, Yuebin Wang, and Jiwen Lu, '3DCNN-DQN-RNN: A Deep Reinforcement Learning Framework for Semantic Parsing of Large-Scale 3D Point Clouds', in *IEEE International Conference on Computer Vision (ICCV)*, pp. 5679–5688, Venice, Italy, (oct 2017). IEEE.

[4] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter, 'Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds', in *Proceedings of the Conference on Computer Vision and Pattern Recognition - CVPR '13*, Portland, Oregon, USA, (2013). IEEE.

[5] Charles R. Qi, Hao Su, Mo Kaichun, and Leonidas J. Guibas, 'PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation', in *Proceedings of the Conference on Computer Vision and Pattern Recognition - CVPR '17*, Honolulu, Hawaii, (2017). IEEE.

[6] Xia Zhong, Mario Amrehn, Nishant Ravikumar, Shuqing Chen, Norbert Strobel, Annette Birkhold, Markus Kowarschik, Rebecca Fahrig, and Andreas Maier, 'Action Learning for 3D Point Cloud Based Organ Segmentation', *Computing Research Repository (CoRR)*, (jun 2018).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 21–22, Santiago de Compostela, Spain, August 29-30, 2020.

Improving Agent Interaction Safety Through Environmental Control and Regulation

Norah Aldahash[1]

Abstract. Safety control of autonomous agent has emerged to be an important topic for research, and much effort is made to provide guarantees on predictable and robust agents. However, a trait of autonomy of agents is to have the ability to interact and exist among other agents, thus increasing the issue of safety. Non-Stationarity of the environment, conflicting goals and heterogeneity of agents are some of the challenges that face safe and optimal behavior of autonomous agents. This study aims to address this problem from the system level, where the environment takes a supervisory role and improving safety of interactions through a governing agent. The role of the governing agent is to monitor and control: run-time monitoring of interaction of agents modeled with an Interaction Petri net (IPN) and controlling environment parameters based on analysis of IPN safety properties.

1 INTRODUCTION

Safe and robust AI has become a major concern as intelligent systems are prevalent now more than ever. There is general consensus on the need to develop techniques for verification and safety control of modern AI systems. The autonomy of agents and their ability to independently make decisions expose risk and impose many challenges. Therefore, there is uncertainty in AI behavior that poses robustness and predictability issues, which are further pronounced and more complex in the instance of multi-agent coordination and cooperation.

Currently there is a notable increase of AI systems that involve multiple agents interacting and coordinating to accomplish tasks, such as disaster area rescue and robotics. Safety issues relating to interaction scenarios that might be harmful to other agents should be assessed rather than only safety issues that exist on the single agent level [4]. Numerous studies have been undertaken to investigate safety problems of autonomous agents but there is a need to further analyze risks of interaction of multi-agent systems.

Safety concerns for autonomous agents operating in a multi-agent environment would not only arise from individual design of agents but may result from interaction and coordination mistakes. Multi agents typically exist in a dynamic, asynchronous, partially observable environments which may compromise correct sequence of message interactions. Moreover, as each agent takes an action towards its goal, it alters its surroundings thus presenting a non-stationary environment. This research approach to resolve such problems arising from interaction of agents through supervisory control.

Methods of verification for multi-agent systems vary but the majority of work focuses on model checking and formal methods during design time. However, such methods need to be complemented with run-time verification of agent interactions. [1]. Safety control of autonomous multi-agent systems can be realized through an external governing agent for monitoring and control. By observing autonomous interaction among agents, a governing agent is designed to identify areas of risk in multi-agent interaction and intervene by altering defined environment parameters to reduce interaction mistakes. While formal methods are based on proving the correctness of system properties run-time monitoring and control is based on continually evaluating a trace of execution and detecting faults. For this purpose, we explore a Colored Petri net model to capture the trace of interaction between agents and propose the Interaction Petri net to be evaluated during system execution for violations of interaction protocols.

This research mimics the Dedale environment [7] designed to create a real-life testbed that is open, dynamic and asynchronous for MAS research. We create a similar environment to evaluate the proposed Petri net model for run-time monitoring and control. Dedale presents a treasure-hunting problem where heterogeneous agents move through a connected graph to collect treasure. The graph also contains Well nodes that cause an agent to terminate when visited. There are different kinds of agents: Collectors to pick up treasure, Explorers to explore the graph for treasure and Tankers who are responsible for accumulating it. However, limited perception and communication range imposes difficulties for agents to interact and achieve their goals.

2 BACKGROUND

There has been a proliferation of work on Petri net based modeling of multi-agent systems for purposes of design, verification and validation. This indicates conformance of Petri nets ability to model multi-agent systems and represent its characteristics as a dynamic, concurrent system.

A Petri net is a graphical model consisting of places represented by circles, transitions represented by bars and arcs which can be directed from places to transitions or from transitions to places. A place can be marked with one token or more and the distribution of tokens among places is called a marking of a net [9]. The net structure is determined by arrangement of places and transitions while movement of tokens between places caused by firing a transition demonstrate its dynamic nature. Previous Petri net models have been created for verification during two different phases of a system life-cycle: design time and run-time. In the analysis and design of MAS Petri nets has shown to naturally model agent architecture and proven to produce feasible models of both agent architecture [3] and agent communication and interaction [5]. Petri net based analysis techniques and simulation facilitated early detection of faults and assessment of liveness and

[1] University of York, UK, email: na1005@york.ac.uk

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 23–24, Santiago de Compostela, Spain, August 29-30, 2020.

dead-lock properties.

On the other hand, run-time models enable verification of system properties. In [2] they focus on temporal properties of a real-time system and present a run-time verification method by comparing observed behaviour with a Time Basic Petri net model of expected behavior and specified time constraints. In contrast to modeling a correct specification of agents behavior, this work aims to create a model which captures run-time behavior for the purpose of analysis and fault detection of interactions.

3 AGENT INTERACTIONS PETRI NET

The goal of the proposed Interaction Petri net is to trace message exchanges between agents for the objective of monitoring and control. Previous work by [5] builds a net to model a specific interaction protocol, alternatively the IPN would model the collection of interactions. The objective is to provide an all-inclusive view to enable the analysis of the state of communication such as: status of send messages, exchange of environment information, open requests. IPN is based on Colored Petri nets, a high-level Petri net that differs from low-level Petri nets in terms of tokens, here tokens hold a data value and belong to types. A place in the IPN represents an agent, a transition represents the action of sending and responding to messages. Tokens are defined to feature messages available to agents with a structure that supports FIPA message structure and metadata [6]. A token is a message instance which the agent intends to send to another agent. The event of firing a transition results in consuming a token from an input place and producing it in an output place. Hence, movement of tokens replicate the dynamic interaction taking place between agents. Based on [8] we define the IPN as a seven-tuple = (P,T,A, Σ,C,G,E):

1. P is finite set of places. A place p is added for each agent.
2. T is a finite set of transitions t such that $P \cap T = \phi$. A transition t is created for each pair of places to model the event of send message.
3. $A \subseteq P \times T \cup T \times P$ is a set of directed arcs. The set of directed arcs represent a link between an agent *place* and the event of send message *transition* .
4. Σ is a finite set of non-empty colour sets. In the IPN a message colour set is defined as a union of data types and holds variables such as performative,sender and receiver.
5. $C : P \rightarrow \Sigma$ is a colour set function that assigns a colour set to each place.
6. $G : T \rightarrow EXPR$ is a guard function that assigns a guard to each transition t which holds a Boolean expression. They define constraints on firing a transition which is based on the interaction protocol.
7. $E : A \rightarrow EXPR$ is an arc expression function that assigns an arc expression to each arc a such that it holds the colour set as the place connected to the arc.

4 ENVIRONMENTAL CONTROL AND REGULATION

Complexity of a multi-agent system require safety control of not only the agent level but the system level as well. We aim to achieve this by introducing parameters of the system environment that are governable to improve the safety of interaction. For example in the described Dedale environment a parameter would be the number of agents: a new agent can be added to replace a terminated agent. An architecture of the monitor and control framework is illustrated in figure 1: it entails capturing a trace of execution through the Interaction Petri net, which then is analyzed by a governing agent to determine if safety properties are satisfied. The result of evaluation determines if it is required to alter environment parameters to ensure correct interaction among agents.

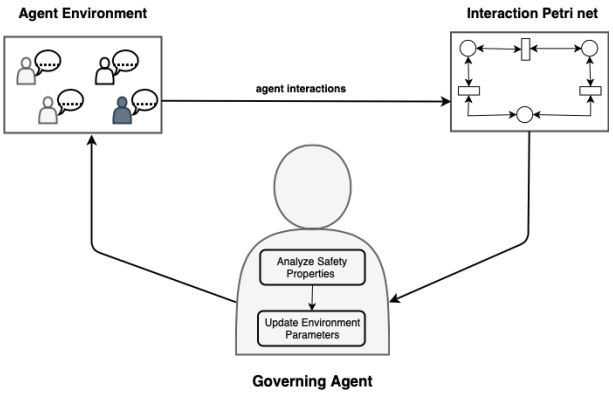

Figure 1. Run-time monitor and control architecture

5 EXPECTED CONTRIBUTION AND CONCLUSION

This research will create a governing agent that monitors and control agent interaction to minimize mistakes. This agent will control defined environment parameters based on observing the proposed Interaction Petri net. A case study has been implemented to evaluate the proposed framework and assess agent interactions with and without a governing agent.

REFERENCES

[1] Najwa Abu Bakar and Ali Selamat, 'Agent systems verification: systematic literature review and mapping', *Applied Intelligence*, **48**(5), 1251–1274, (2018).

[2] Matteo Camilli, Angelo Gargantini, Patrizia Scandurra, and Carlo Bellettini, 'Event-based runtime verification of temporal properties using time basic petri nets', in *NASA Formal Methods Symposium*, pp. 115–130. Springer, (2017).

[3] Jose R Celaya, Alan A Desrochers, and Robert J Graves, 'Modeling and analysis of multi-agent systems using petri nets', in *2007 IEEE International Conference on Systems, Man and Cybernetics*, pp. 1439–1444. IEEE, (2007).

[4] Nader Chmait, David L Dowe, David G Green, and Yuan-Fang Li, 'Agent coordination and potential risks: Meaningful environments for evaluating multiagent systems', in *Evaluating General-Purpose AI, IJCAI Workshop*, (2017).

[5] R Scott Cost, Ye Chen, Tim Finin, Yannis K Labrou, Yun Peng, et al., 'Modeling agent conversations with colored petri nets', in *Working notes of the Autonomous Agents' 99 Workshop on Specifying and Implementing Conversation Policies*, (1999).

[6] ACL Fipa, 'Fipa acl message structure specification', *Foundation for Intelligent Physical Agents, http://www. fipa. org/specs/fipa00061/SC00061G. html (30.6. 2004)*, (2002).

[7] Cédric Herpson, 'Dedale: A dedicated testbed for multi-agents problems', (2019).

[8] Kurt Jensen and Lars M Kristensen, *Coloured Petri nets: modelling and validation of concurrent systems*, Springer Science & Business Media, 2009.

[9] Tadao Murata, 'Petri nets: Properties, analysis and applications', *Proceedings of the IEEE*, **77**(4), 541–580, (1989).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 23–24, Santiago de Compostela, Spain, August 29-30, 2020.

Recognition and learning of agent behaviour models

Stanislav Sitanskiy[1]

Abstract. This paper provides an overview of my PhD research project which focuses on the definition of a behaviour model as a strategy that an agent uses on selecting actions for solving problems, and on Behaviour Trees as a tool for describing the specified model. The work, firstly, aims to finding the best method for recognizing a behavior model in a given action plan o data received from sensors. This will be followed by developing the framework for behaviour learning from the same type of data and its representation by means of Behaviour Trees. The findings can contribute significantly to video game designers and developers, smart-things area, digital assistants and guides, robotics, etc.

1 INTRODUCTION

Research on AI is more and more focusing towards explainable technology that accounts for the outcomes of programs and products. One important aspect in this direction is the ability to recognize the behaviour patterns of an application in order to make sensible and informed decisions, predictions or explaining selected choices. Another equally important aspect is the possibility of learning agents behavior to accelerate the construction of intelligent behaviours, or the investigation of certain fragments of the learned behaviour. The behaviour of a system goes beyond the model that governs the physics of the domain; i.e., the model determines the legal actions that can be performed and under which conditions whereas the behaviour determines the reasons for choosing one action among many options or the order in which actions will be performed. Generally speaking, the model tells us what can be done, the behaviour tell us how this is done.

Differences in behavioral models can be illustrated by the examples of plans in Table 1, executed by two planning agents when solving a problem of a *Logistics* domain, a classical planning domain introduced in the first International Planning Competition[2]. This problem consists in transporting packages `P1` and `P2` from location `L1` to location `L2` with two available trucks, `T1` and `T2`. Plan A follows Behaviour 1 (which consists in loading all of the packages in the truck before starting to unload), whereas plan B follows Behaviour 2 (which consists in transporting all the packages one by one).

There exist various methods that can be exploited for describing the behaviour of an agent. Finite State Machines (FSMs) are widely used, for instance, for decision making in game AI. The difficulty in reusing transitions in FSMs is a major limitation in planning since addressing the same goal for a number of objects of the same type requires using the same type of actions. Another solution is using Behaviour Trees (BTs), which focus on making individual states modular and so they can be easily reused in different parts of the behaviour[8, 3, 2]. A BT is a static, tree-like, hierarchical structure representing the behaviour of an agent in a human-editable way. It has origin in the video-game industry and was introduced to reduce games AI-definition complexity and substitute FSMs. It is actively used to model the behaviour of non-player characters (NPCs) in video-games and have been used in Halo, Bioshock, and Spore. This technology quickly gained popularity and now it is included in the major game engines like CryEngine, Unreal Engine, and Unity. Usually BTs are created by hand by game designers and developers to define the agent action policy and parts of them can be mixed or reused to create new behaviours for other agents.

In our work, we aim to uncover the behaviour of a planning agent and develop a system which, on the base of the physical model of the world, could automatically extract behaviour from data received from sensors or user activity logs to create the BT that describes it. We have divided the achievement of this goal into two stages: first, we will focus on the agent behaviour recognition and, then, we will work on methods for learning this behaviour.

Table 1. Two example plans, representing different behaviours: Plan A - load all of the packages in the truck before starting to unload and Plan B - transport all the packages one by one

Plan A	Plan B
A1: LOAD P1 T1 L1	A1: LOAD P1 T1 L1
A2: LOAD P2 T1 L1	A2: DRIVE T1 L1 L2
A3: DRIVE T1 L1 L2	A3: UNLOAD P1 T1 L2
A4: UNLOAD P1 T1 L2	A4: DRIVE T1 L2 L1
A5: UNLOAD P2 T1 L2	A5: LOAD P2 T1 L1
	A6: DRIVE T1 L1 L2
	A7: UNLOAD P2 T1 L2

2 STATE OF THE ART

Discovering behaviour of a planning agent is closely related to the concept of activity and plan recognition [14] but it stresses a different dimension of the acting agent. While plan recognition puts the focus on identifying the goal or intention of the agent, and the corresponding path to achieve it, given some observations of its behaviour [9, 4, 13], a behavioural analysis aims at finding the patterns that govern the know-how of the agent. In this sense, it can be argued that analyzing the behaviour of a planning agent resembles the purpose pursued within pattern discovery in data mining [6]. The objective is certainly similar but a key distinction lies in that a planning agent builds up a plan by using an action model and so the behaviour of the agent does not solely respond to some goal-driven strategy but also to some (unknown) model that governs the physics of the domain.

[1] Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Valencia - Spain, email: stasiig@inf.upv.es

[2] http://ipc98.icaps-conference.org/

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 25–26, Santiago de Compostela, Spain, August 29-30, 2020.

As far as we know, most of the research in this area is focused on activity and plan recognition.

With respect to the problem of agent behaviour learning, several approaches for learning this behaviour in the form of BTs can be found. In [1], a method based on Reinforcement Learning (RL) is described for learning BTs in unknown dynamic environments. The BT is learned through the combination of a greedy algorithm with genetic programming (GP). However, due to the GP, this approach requires good computational power to have a result in reasonable time and it needs the environment prepared for RL, which is not always available in real-life applications.

Another work[11] presents *Learning by Observation* methods for the construction of BTs that will be used in the high level control of the widely known real-time strategy game Star Craft. In a nutshell, this method consists of creating a maximally specific tree based on the player's actions log, which iteratively is reduced due to the searching for identical patterns using motif finding (GLAM2 software[5]). These patterns are then combined into new sequences and inserted into the original tree. The process ends when all patterns are found. Although proved to reduce sequences of player actions by almost two orders of magnitude, this approach still produces large BTs which makes difficult to them be edited by hand and it is not sophisticated enough to separate out most parallel or reactive actions. Moreover, the way in which the trees are reduced necessarily removes information, so it is possible that important information is lost in the process.

3 BEHAVIOUR RECOGNITION

In accordance with the first part of the thesis, we are interested to find out the strategy that the agent uses when calculating the plans that solve the planning problems. In order to understand the behaviour of the planning agent, we first putted the focus on the order of solving the problem goals. Particularly, we were interested in studying if the agent follows some behavioural pattern to address the problem; e.g., sequentially solving the problem goals, interleaving actions oriented at achieving different goals or any combination of these strategies.

Therefore, we used BTs to define several behaviours for solving the problems of the Logistics domain [12]. Then, the planner LAMA[10] was used to solve the same problems with different versions of the original domain, adapted to simulate these behaviours. Finally, each problem solved with LAMA was compared to the plans for that problem obtained with the BTs in order to decide which behaviour was following. For doing this, we tested plan-plan distance metrics presented in work[7]. In addition, we introduced a novel metric, which is an extension of the metric based on sequence of states that uses disjunctive landmarks. Our experiments show that using these metrics to recognize the underlying behavior has some limitations. Mainly, these metrics, based in plan-plan distance, have difficulties in identifying a behavior when both plans, even following the same behavior, differ in the order in which the agent reaches the goal, as well as in the method of choosing the resources to achieve the goal.

Trying to overcome these limitations, we are currently working on the improvement of recognition methods by using machine-learning algorithms. Specifically, we are working on splitting plans on n-grams of a given length before applying a classifier (Naive Bayes, Support Vector Machine, Decision Tree). N-gram usage allows to extract the action order and resource usage in the action context, or to detect the object evolution. An additional technique that we are testing is to preprocess the plan for better feature extraction. This preprocess consists in resource anonimization, allowing to resolve the syntactical difference between semantically equivalent plans.

4 BEHAVIOUR LEARNING

Upon completion of the previous stage and the best way for recognizing a behavior model is selected, we plan to proceed to the development of a methodology for behavior learning. If the goals are successfully achieved, it will be provided a technique that will greatly simplifies the creation, modification and maintenance of behavior models for agents that can be used in various areas of IT, for example, controlling characters in computer games, digital assistants and guides, smart-things, robotics, etc.

ACKNOWLEDGMENTS

This work is supported by the Spanish MINECO project TIN2017-88476-C2-1-R and the FPI grant PRE2018-083896.

REFERENCES

[1] M. Colledanchise, R. Parasuraman, and P. Ögren, 'Learning of behavior trees for autonomous agents', *IEEE Transactions on Games*, **11**(2), 183–189, (2019).

[2] Michele Colledanchise, Diogo Almeida, and Petter Ögren, 'Towards blended reactive planning and acting using behavior trees', in *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pp. 8839–8845, (2019).

[3] Michele Colledanchise and Petter Ögren, *Behavior trees in robotics and AI: An introduction*, CRC Press, 2018.

[4] Richard G. Freedman, Hee-Tae Jung, and Shlomo Zilberstein, 'Plan and activity recognition from a topic modeling perspective', in *24th International Conference on Automated Planning and Scheduling, ICAPS*, (2014).

[5] Martin C. Frith, Neil F. W. Saunders, Bostjan Kobe, and Timothy L. Bailey, 'Discovering sequence motifs with arbitrary insertions and deletions', *PLoS Computational Biology*, **4**(5), e1000071, (2008).

[6] Wensheng Gan, Jerry Chun-Wei Lin, Philippe Fournier-Viger, Han-Chieh Chao, and Philip S. Yu, 'A survey of parallel sequential pattern mining', *ACM Trans. Knowl. Discov. Data*, **13**(3), 25:1–25:34, (2019).

[7] Tuan Anh Nguyen, Minh Do, Alfonso Emilio Gerevini, Ivan Serina, Biplav Srivastava, and Subbarao Kambhampati, 'Generating diverse plans to handle unknown and partially known user preferences', *Artificial Intelligence*, **190**, 1–31, (2012).

[8] Ricardo Palma, Pedro A. González-Calero, Marco Antonio Gómez-Martín, and Pedro Pablo Gómez-Martín, 'Extending case-based planning with behavior trees', in *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, May 18-20, 2011, Palm Beach, Florida, USA*, (2011).

[9] Miquel Ramírez and Hector Geffner, 'Plan Recognition as Planning', in *21th International Joint conference on Artifical Intelligence, IJCAI*, pp. 1778–1783. AAAI Press, (2009).

[10] Silvia Richter and Matthias Westphal, 'The lama planner: Guiding cost-based anytime planning with landmarks', *Journal of Artificial Intelligence Research*, **39**, 127–177, (2010).

[11] Glen Robertson and Ian D. Watson, 'Building behavior trees from observations in real-time strategy games', *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, 1–7, (2015).

[12] S. Sitanskiy, L. Sebastia, and E. Onaindia., 'Behaviour recognition of planning agents using behaviour trees', in *Proceedings of the 24th International Conference on Knowledge-Based and Intelligent Information Engineering Systems*, (2020(in press)).

[13] Shirin Sohrabi, Anton V. Riabov, and Octavian Udrea, 'Plan recognition as planning revisited', in *25th International Joint Conference on Artificial Intelligence, IJCAI*, pp. 3258–3264, (2016).

[14] Gita Sukthankar, Robert P. Goldman, Christopher Geib, David V. Pynadath, and Hung Bui, *Plan, Activity, and Intent Recognition: Theory and Practice*, Morgan Kaufmann, 2014.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 25–26, Santiago de Compostela, Spain, August 29-30, 2020.

Integrating Open-ended Learning and Planning for Long-Term Autonomy

Gabriele Sartor[1]

Abstract. Classical planning is still a powerful tool able to perform rather complex reasoning on domains defined by a high-level representation. However, its main problem is the lack of flexibility in the definition of the domain. Once the representation of the world is defined by the expert, the capabilities of the agent are fixed and, consequently, also its potentially achievable goals. For this reason, many researchers have shifted their attention on developing systems able to produce autonomously a high-level representation of the world, resulting from the experience gathered during the interaction with the surrounding environment. IMPACT (Intrinsically Motivated Planning Architecture Curiosity-driven roboTs)[2] has been our first attempt to implement a software architecture using high-level planning and able to extend its operational capabilities.

1 INTRODUCTION

In the last decades, traditional planning robotic systems have been developed providing the agent the knowledge necessary to perform tasks defined at design time. However, in some cases, particularly in space exploration missions [6], the agent could need to deal with unforeseeable situations or simply detect a variation in the environment dynamics without the human intervention.

Consequently, many researchers have started to study new methods to abstract knowledge learned during the interaction with real world in order to encapsulate the complexity under an easier representation, suitable for performing high-level planning. Recently, some methods have been proposed to translate the agent's experience in PDDL representation [7]. For example, it has been proposed an algorithm translating low-level information about the initial and final states of each action, called options, into a fully working propositional symbolic planning domain [5]. Some works have also started to reconcile deep learning with more abstract representations [4]. For instance, a system demonstrated the possibility to autonomously generate a first-order logic (FOL) representation compatible with symbolic classical planning, using neural networks [1]. In this case, the most important element developed has been a particular autoencoder, able to transform the feature vectors of the objects visualized in images into a FOL description, and vice versa.

The project IMPACT [9, 8] has extended the previous work creating an open-ended learning system [3] able to learn new capabilities interacting with a simulated environment, abstract this knowledge into propositional PDDL as in [5] and plan on it.

[1] Università degli Studi di Torino, Italy, email: gabriele.sartor@unito.it

[2] This research has been supported by the European Space Agency (ESA) under contract No. 4000124068/18/NL/CRS, project IMPACT (Intrinsically Motivated Planning Architecture for Curiosity-driven roboTs). The view expressed in this paper can in no way be taken to reflect the official opinion of the European Space Agency.

2 CURRENT RESEARCH

In the last years, the importance of space missions is increasing. In particular, the next most challenging missions are focusing on the exploration of Mars. The robots designated to perform the experiments on-site will have to be equipped with all the knowledge necessary to deal with their duties with a high degree of autonomy because of the latency of communications between the Earth and Mars. For this reason, it is important to build systems as autonomous as possible, but also able to extend their operational capabilities on-site in order to face changes in the environment, unforeseeable events and increase the utility of the mission discovering new aspects of the world and achieve goals unknown at design time. IMPACT has been our first try to develop such software architecture.

This system is implemented as a three-layered architecture integrating the following main modules:

1. *Planning*, reasoning on the initial knowledge of the world and operational capabilities included in the agent at design time used to reach the mission goals;
2. *Abstraction*, translating the low-level data gathered during the mission into a propositional domain of the environment;
3. *Learning*, responsible for learning new skills, triggered by the robot's curiosity [10].

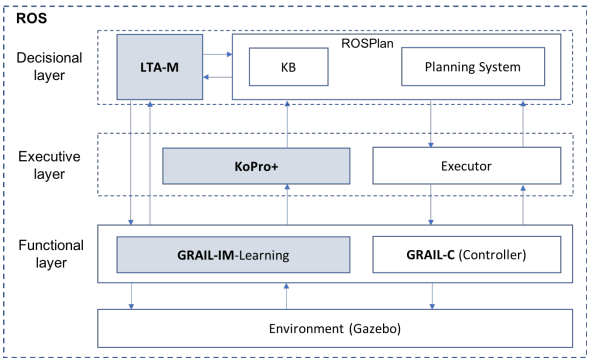

Figure 1. IMPACT high-level software architecture.

In figure 1 it is shown the architecture of our software system, divided in three layers. The lower one, *the Functional layer*, contains the modules dealing with the low-level information of the environment. It implements the controller of the robot and the learning ability integrating the GRAIL system [10], an intrinsically motivated reinforcement learning (IM-RL) component able to extend the competences of the agent. This component can autonomously discover interesting states of the environment and, using a competence-based

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 27–28, Santiago de Compostela, Spain, August 29-30, 2020.

IM-RL algorithm, acquires a new skill to reach that significant configuration.

Instead, the *Executive layer* is a mapping level in which (i) the high-level planning operators are translated into commands for the controllers and (ii) low-level information is transformed into a PDDL representation of the world [5], suitable for planning. In particular, the abstraction of the low-level information is crucial for such systems, because it defines the type of representation and, consequently, the limitations of the higher level. For example, despite its simplicity, the propositional representation can limit the expressiveness of the higher level, given that it is not possible to generalize similar actions and features.

The high-level planning domain and the planning system, implemented using the ROSPlan framework [2], are included inside the *Decisional layer*. This level is responsible for deciding how the high-level goals have to be achieved and controlling the execution feedbacks received from the lower layers, in order to potentially replan its activities. The top layer has also to decide the schedule of the goals to be achieved. For this purpose, it has been developed also a simple component called *Long-Term-Autonomy Module* (LTA-M). The idea of such component is to create a decisional module selecting the next tasks to be performed in the future, in order to increase as much as possible its satisfaction in terms of mission goals and extension of operational capabilities. In other words, it implements strategies to alternate the achievement of intrinsic and extrinsic goals.

The system described in this paper has been tested in two different space exploration scenarios. In particular, we simulated a scenario in which a robotic arm is sent to Mars in order to pickup some stone samples necessary for its experiments [8]. In the simulation, we assume that the robot is designated to explore two interesting valleys in which the scientists of the mission control center want to perform analysis of the ground. In the first valley, the agent is able to pickup samples of stones without any difficulty, while in the second one the stones have a particular concave shape. During its activities, the agent tries to pickup a vase-shaped stone in the second valley, discovering that it is not able to grasp with its current capabilities because of its dimensions, failing to reach its high-level goal. The failure triggers its curiosity towards the unexpected situation and tries to learn a new way to grasp the unknown object. The intrinsic motivated component GRAIL [10] detects a curious state and stores it as intrinsic goal. When the agent decides to focus on that interesting state, it will try to reach again that situation with a high level of confidence. After several attempts, the system learns to grasp the new object and, exploiting the execution layer, to abstract this new discovered aspect of world creating a new symbol and operator to deal with it. In this way, it has been demonstrated that the architecture has been able to extend its operational capabilities and the potential of this system.

3 FUTURE WORKS

The architecture presented in the paper could be still improved in different aspects. One of the most important limitations of the system is the use of a high-level planning propositional representation. This approach is not effective in terms of scalability, because each aspect of the world has to be represented with a specific symbol. Instead, first-order logic representation can use parametrized symbols and operators to generalize some aspects in a more compact description. Given the potential of the autoencoders in generating a representation from the world [4], we will examine the possibility of combining them with open-ended learning to deal with such limitation.

Another component to be examined in depth is the LTA-M [6], which still presents a simple form. We will examine the possible strategies to manage the long-term needs of the agent based on the goal required by the user and its curiosity. A possible extension could be exploiting the high-level planning to reach less explored states in order to increase the possibility of encountering new intrinsic goals and, consequently, learning new aspects of the world. Lastly, our intention is to generalize this architecture in order to extend the range of applicability of this methodology.

4 CONCLUSION

We proposed an open-ended learning architecture able to learn new competences, abstract this knowledge and reasoning over this new representation extending its operational capabilities. Robotics system autonomy is a branch of research of increasing importance. However, in this field, autonomy and interpretability seem to be two orthogonal aspects[1]. We think that an abstract representation of the domain will be fundamental for keeping track of the agent's improvement caused by its learning components. Consequently, we intend to improve the robustness of this architecture and generalize it for a wider range of applications.

ACKNOWLEDGEMENTS

For more details on the system implementation, please see the main conference article *"Integrating open-ended learning in the sense-plan-act robot control paradigm"* [8].

REFERENCES

[1] Masataro Asai, 'Unsupervised grounding of plannable first-order logic representation from images', in *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pp. 583–591, (2019).

[2] M. Cashmore, M. Fox, D. Long, D. Magazzeni, B. Ridder, A. Carrera, N. Palomeras, N. Hurtos, and M. Carreras, 'Rosplan: Planning in the robot operating system', in *Twenty-Fifth International Conference on Automated Planning and Scheduling*, (2015).

[3] S. Doncieux, D. Filliat, N. Díaz-Rodríguez, T. Hospedales, R. Duro, A. Coninx, D. M. Roijers, B. Girard, N. Perrin, and O. Sigaud, 'Open-ended learning: a conceptual framework based on representational redescription', *Frontiers in neurorobotics*, **12**, 59, (2018).

[4] M. Garnelo and M. Shanahan, 'Reconciling deep learning with symbolic artificial intelligence: representing objects and relations', *Current Opinion in Behavioral Sciences*, **29**, 17 – 23, (2019). SI: 29: Artificial Intelligence (2019).

[5] G. Konidaris, L. P. Kaelbling, and T. Lozano-Perez, 'From skills to symbols: Learning symbolic representations for abstract high-level planning', *Journal of Artificial Intelligence Research*, **61**, 215–289, (2018).

[6] L. Kunze, N. Hawes, T. Duckett, M. Hanheide, and T. Krajník, 'Artificial intelligence for long-term robot autonomy: A survey', *IEEE Robotics and Automation Letters*, **3**(4), 4023–4030, (2018).

[7] D. Mcdermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins, 'PDDL - The Planning Domain Definition Language', Technical report, CVC TR-98-003/DCS TR-1165, Yale Center for Computational Vision and Control, (1998).

[8] A. Oddi, R. Rasconi, V.G. Santucci, G. Sartor, E. Cartoni, F. Mannella, and G. Baldassarre, 'Integrating open-ended learning in the sense-plan-act robot control paradigm', *24th European Conference on Artificial Intelligence, ECAI 2020*, (2020).

[9] A. Oddi, R. Rasconi, V.G. Santucci, G. Sartor, E. Cartoni, F. Mannella, and G. Baldassarre, 'An intrinsically motivated planning architecture for curiosity-driven robots', *CEUR Workshop Proceedings*, **2594**, 19–24, (2020).

[10] V. G. Santucci, G. Baldassarre, and M. Mirolli, 'Grail: a goal-discovering robotic architecture for intrinsically-motivated learning', *IEEE Transactions on Cognitive and Developmental Systems*, **8**(3), 214–231, (2016).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 27–28, Santiago de Compostela, Spain, August 29-30, 2020.

A Computational Trust Model for Task Delegation

Jeferson J. Baqueta [1]

Abstract. In cooperative environments is common that agents delegate task to each other to achieve their goals, since an agent may not have the capabilities or resources to achieve its objectives alone. Thus, in order to delegate a task, the agent needs to deal with information about the abilities, expertise, and goals of their partners. However, the lack or inaccuracy of these information may affect the agent's judgment about a given partner; and hence, increases the risks to rely on an untrustworthy agent. Therefore, in this work, we aim to present a trust model that combines social image and reputation concepts in order to produce a most reliable information about the characteristics and abilities of agents.

1 INTRODUCTION

In Multi-agent systems (MAS), trust and reputation models have been adopted as an important solution in order to ensure security and efficiency [2][8]. In particular, these models offer mechanisms to punish inappropriate behaviors of agents and improve the partner selection process in uncertain situations [4]. Generally, in this kind of models, the agents are rated (*e.g.*, good or bad) as they interact with other agents in the society. Thus, a good partner may be identified based on the ratings about its behaviour. In turn, these ratings, are performed along the time considering a certain context (*i.e.*, the agent's characteristics such as velocity, strength, or expertise).

In task delegation scenarios, evaluating partner agents just considering their characteristics (internal factors) may be risky, since delegation involves to rely on someone [3]. Therefore, the environment conditions, as well the risks associated to a given decision (external factors), should be considered before delegate a task to a partner. In this sense, in [5] a trust model that combines the internal and external factors into a trust measure is proposed, which is adopted in [3] to evaluate the risks to delegate tasks. In particular, in the literature there are other works and models that use trust to solve the task delegation problem (*e.g.*, [6][1]). However, in general, the trust concept is modeled based on the agents' capabilities or the task requirements (*e.g.*, cost, time, and quality) (internal factors).

In this work, we aim to extend the trust model proposed in [5]. In our approach, the trust measure is computed through the social image (agent's opinion) [4], reputation (third party's opinion) [9], and agents' goals. In [5], it is assumed that the inputs of the model are fed by the beliefs of agents, which are produced through agents' direct experiences. In particular, the use of image and reputation allows that the agents share their experiences each other, increasing the volume of information available. Thus, the increase of the number of information sources tends to improve the decision-making process of agents, including the decisions about delegation. Moreover, in the model presented in [5], some parameters, such as the causality relationship among the concepts of the model, have their values predefined and need to be configured by a human specialist. In particular, this configuration is essential according to the application, since the agents' profiles rely on this context. In our approach, we believe that these configuration could be defined dynamically by agents themself according to their goals. Since this research is its initial state, we have focused on the use of social image and reputation concepts to compute the internal factors elements discussed in the trust model.

[1] Federal University of Technology - Parana (UTFPR), Brazil, email: jefersonbaqueta@gmail.com

2 TRUST MODEL STRUCTURE

In our trust model, we consider that the task delegation is based on the skills of the agent partner. In particular, an agent may have several skills and each skill can be decomposed in different criteria. Thus, an agent may receive different ratings about its skills according to the needs and expectations of the appraiser agent. In this case, a given agent can play more than one role in the society and receive a distinct evaluation for each one of them, including, being able to stand out more in one role. For instance, a doctor (skill) can be evaluated considering several criteria, such as his experience, service price, among others. At the same time, this doctor could be evaluated by his skill to play the guitar. In this case, a same agent plays two distinct roles, which must be evaluated considering different criteria. Therefore, an appraiser agent could evaluate the target agent based on the target's reputation and/or its personal image about the target's skills.

As discussed in [4], the social image is a social evaluation asso-

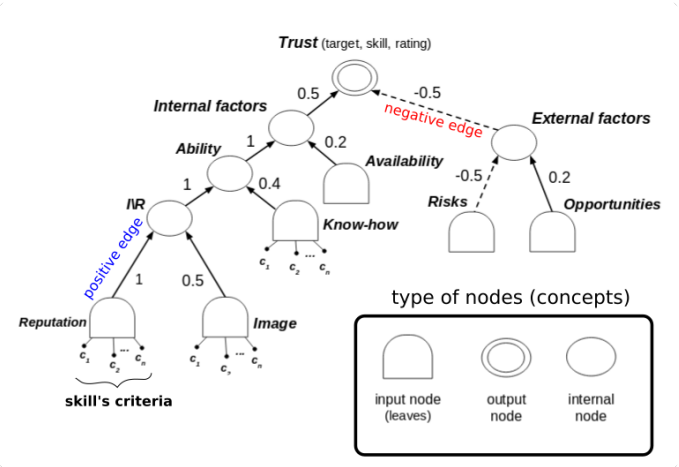

Figure 1. The proposed trust model and the relationship among concepts (Fuzzy Cognitive Map).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 29–30, Santiago de Compostela, Spain, August 29-30, 2020.

ciated to functions or roles that a given target plays on. Generally, the social image is associated to a personal appraiser's opinion about a target. In turn, reputation can be defined as an opinion shared by the majority of the member of the society, which can be built based on gossip. Moreover, as discussed in [8], image and reputation can have different relevance for the appraiser agent. In this sense, the personal opinion of an agent may be more important and significant for it than the opinion from other agents, and vice-versa. This situation is represented in Figure 1, where the image affects more significantly the I/R node than reputation. Therefore, in our model, the opinion of third parties can be compared to personal opinion of an agent during the decision process of rely on another agent.

Our trust model is built using a Fuzzy Cognitive Map (FCM) [7], as it shown in Figure 1. A FCM is defined through a set of nodes connected by weighted edges. The nodes represent the concepts that describe the evaluation characteristics, and the weighted edges represent the causal relationships that exist among concepts [10]. Generally, the values assigned to edges are represented by the interval [-1,1], where -1 means a negative causal relationship, 0 means neutral, and 1 means a positive relationship. Thus, when two concepts are connected by a positive edge, the increase on the value of one of concepts causes an increase of the value of other concept (positive causality). On the other hand, when the concepts are connected by a negative edge, the increase on the value of one of concepts causes a decrease on the value of other concept (negative causality).

Furthermore, in our approach, an appraiser agent may generate a different FCM for each target's skill, where each FCM is represented as a tree. In particular, the leaf nodes represent the criteria associated to the evaluated skill and the others nodes represent the essential concepts of the model. As it is possible to see in Figure 1, to estimate the value of internal factors node, in our evaluation, we consider the social image and reputation of a given target agent, the target's know-how, which represents the previous services provided by the target for other agents (*i.e.*, target's references), and the target's availability, which indicates if the target is available to perform the task in a given instant of the time. In relation to external factors, the risks and opportunities can be estimated based on the beliefs and perceptions of agents about the environment (*e.g.*, privileged information or an obstacle that may prevent the execution of a given task).

It is important to note that an appraiser agent computes the social image about a given target agent aggregating its impressions produced from the past interactions with this agent (its personal experiences). In particular, an impression is a rating that an appraiser agent performs about the service provided by a target agent (after the target completes the delegated task). In turn, the aggregation of impressions into a social a image is performed through a weighting mean, similar as defined in [9]. Moreover, the appraiser may share its social image with other agents in the society. In this case, the shared image will circulate in the society as a gossip. On the other hand, when an agent receives different images about the same target agent, this agent can aggregate these images into a reputation.

In conclusion, in order to compute a trust measure, the input nodes' values of the FCM are propagate until arrive at the output node. Thus, for a given node *n* in a level *i*, its value is computed taking the sum of the values of all nodes in the level $i-1$ (connected to *n*) multiplied by the corresponding edge weights. In particular, using a threshold function (*e.g.*, $f(x) = tanh(x)$), the resulting node value is squashed into the interval [–1,1], where the value -1 means the worst rating value for a given concept, 0 means a neutral rating, and 1 means the best value.

3 CONCLUSION AND FUTURE WORKS

We have presented a general view about a model of trust that takes into account internal and external factors to compute a trust measure. In this work, we only discuss the computation of internal factors of this trust model, which is based on the social image and reputation information. As future work, we intend to explore means to compute the risk and opportunities that make up the external factors of the trust model. Moreover, as a starting point, we intend to study some of the works discussed in [1], since these works provide approaches to handle the risks (environmental risk).

Additionally, we intend to propose a trust model where the goals of agents are taken into account before they decide to rely on another agent. In this case, the agent's goals could be used as indicators for the risk decisions making, since according to the state and importance of a goal, the agent may or not take a more risky stance. In particular, we intend to evaluate this model based on the amount and quality of the completed tasks performed by partner agents.

Finally, as discussed in [3], the use of FCM depends on the human intervention, since the edges' values (causal relationship among the concepts) must be defined by a specialist. We believe that these weights could be defined by agents themselves according to their goals, since the agents could adjust the causal relationship among the concepts of the FCM as they learn with the environment or with other agents.

ACKNOWLEDGMENTS

This work is fully founded by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). In particular, the author would also like to thank his thesis advisors Prof. César Augusto Tacla and postdoctoral researcher Mariela Morveli-Espinoza for their valuable suggestion and comments.

REFERENCES

[1] Diego De Siqueira Braga, Marco Niemann, Bernd Hellingrath, and Fernando Buarque De Lima Neto, 'Survey on computational trust and reputation models', *ACM Computing Surveys (CSUR)*, **51**(5), 1–40, (2018).

[2] Francesco Buccafurri, Antonello Comi, Gianluca Lax, and Domenico Rosaci, 'Experimenting with certified reputation in a competitive multi-agent scenario', *IEEE Intelligent Systems*, **31**(1), 48–55, (2015).

[3] Christiano Castelfranchi and Rino Falcone, *Trust theory: A socio-cognitive and computational model*, volume 18, John Wiley & Sons, 2010.

[4] Rosaria Conte and Mario Paolucci, *Reputation in artificial societies: Social beliefs for social order*, volume 6, Springer Science & Business Media, 2002.

[5] Rino Falcone, Giovanni Pezzulo, and Cristiano Castelfranchi, 'A fuzzy approach to a belief-based trust computation', in *Workshop on Deception, Fraud and Trust in Agent Societies*, pp. 73–86. Springer, (2002).

[6] Nathan Griffiths, 'Task delegation using experience-based multi-dimensional trust', in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pp. 489–496, (2005).

[7] Bart Kosko et al., 'Fuzzy cognitive maps', *International journal of man-machine studies*, **24**(1), 65–75, (1986).

[8] Isaac Pinyol and Jordi Sabater-Mir, 'Computational trust and reputation models for open multi-agent systems: a review', *Artificial Intelligence Review*, **40**(1), 1–25, (2013).

[9] Jordi Sabater and Carles Sierra, 'Regret: reputation in gregarious societies', in *Proceedings of the fifth international conference on Autonomous agents*, pp. 194–195, (2001).

[10] Chrysostomos D Stylios and Petros P Groumpos, 'Modeling complex systems using fuzzy cognitive maps', *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, **34**(1), 155–162, (2004).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 29–30, Santiago de Compostela, Spain, August 29-30, 2020.

Verification and Validation of Planning-based Autonomous Systems

Anas Shrinah [1]

Abstract. The verification and validation of planning domain models is one of the biggest challenges to the deployment of planning-based automated systems. In this work, we propose an approach that increases the robustness of planning domain model verification by eliminating false positive counterexamples using model checkers and state trajectory constraints planning techniques. Additionally, we define planning domain model functional equivalence and develop a method to compare the functionality of two planning domains by integrating planners and Boolean satisfiability (SAT) solvers.

1 INTRODUCTION

Erroneous planning systems could fail to find a plan, generate unrealistic plans that will fail to execute, produce unsafe plans or improperly execute correct ones. In planning-based automated systems, such flaws could cause catastrophic consequences in real world operations. This project aims to provide automated support to verify planning-based autonomous systems are clear of such errors.

This work seeks to address three challenges: performing robust planning domain model verification of safety properties; validating the correctness of a planning domain model with respect to a given formal requirement in an automated fashion; and automatically debug failed test cases in test-based verification of planning-based autonomous systems to pinpoint the root cause of such failures.

The following sections summarise the work done on the first two challenges.

2 PLANNING DOMAIN MODEL VERIFICATION OF SAFETY PROPERTIES

The state-of-the-art verification methods for planning domain models are susceptible to false positive counterexamples, i.e. counterexamples that are unreachable by sound planners when the domain under verification is used for a planning task. Designers can be misguided to needlessly constraint domain models due to such unrealisable counterexamples, thereby potentially inhibiting necessary behaviours. In addition, false positive counterexamples can lead verification engineers to overlook counterexamples that are reachable by planners.

Our proposed method to overcome this shortcoming is to use planning goals as constraints during verification, a concept borrowed from verification and validation research [3]. Thus, we introduce *goal-constrained planning domain model verification*, an approach that eliminates invalid planning counterexamples per se.

[1] University of Bristol,UK, email: anas.shrinah@bristol.ac.uk

A valid planning counterexample is defined to be a sequence of actions that falsifies a given safety property before achieving the planning goal from the initial state, while none of the sub-sequences of the counterexample can achieve the goal.

Definition 2.1. A **valid planning counterexample** for a safety property, p, of a planning problem is a *non-redundant plan*, π, that falsifies the safety property, $\pi \not\models p$.

Valid counterexamples have to be non-redundant to exclude any plans that are enriched with action sequences that are unnecessary to achieve the planning goal but required to falsify the safety property. Therefore, the scope of our method is limited to non-redundant planners, as sound planners can generate plans that have redundant subsequences.

Thus, to guarantee the resulting counterexamples are valid, we constrain the verification problem with the planning goal, and we exclude any counterexample that is a redundant plan.

2.1 Goal-constrained planning domain model verification using model checkers

To force model checkers to return only valid planning counterexamples, the safety property is first negated and joined with the planning goal in a conjunction. This conjunction is then negated and supplied to the model checker as an input property. The final property requires the model checker to find a counterexample that both, falsifies the safety property and satisfies the planning goal. Unlike Def. 2.1, these steps permit sequences that falsify the property after satisfying the goal. However, planners terminate the search once the goal has been achieved, thus rendering such sequences unreachable. These sequences are prohibited by augmenting the model transitions guards with the negation of the goal. This restricts all transitions once the goal is achieved.

2.2 Goal-constrained planning domain model verification using planning techniques

The state trajectory constraints introduced in PDDL3.0 [4] are used to perform planning domain model verification. Firstly, the negation of the safety property is expressed using PDDL3.0 modal operators and embedded in the original domain model as a state trajectory constraint. Then a planner uses the modified model to perform the verification process. If the modified domain model permits producing plans that, along with achieving the goal, contradict the safety property, then the original model can also produce such unsafe plans. Thus, the returned plan constitutes a valid planning counterexample. On the other hand, if the planner fails to find a plan for the given

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 31–32, Santiago de Compostela, Spain, August 29-30, 2020.

goal while using the modified domain, then the original domain is safe and the safety property holds in any plan produced for that goal.

3 PLANNING DOMAIN MODEL VALIDATION

Planning domain model validation is one of the key aspects of Knowledge Engineering in Planning and Scheduling (KEPS). Among other tasks, this activity is concerned with checking the correctness of a planning domain model with respect to a given reference. If the requirements are described informally, then the process of validating the domain model is also informal [5]. On the other hand, when the requirements are described formally, it is feasible to perform formal validation.

There are applications where a planning domain model and its reference are both given in the same formal language, like PDDL. For example, evaluating the quality of automated planning model learning algorithms. For this application, a hand crafted model is used to generate a number of plans that are fed to the model learning method and the produced planning domain model is validated against the input model [1, 7].

Another application is related to planning domain models optimisation methods [2, 6]. These techniques learn operator sequences, called macro-operators, and add them to the planning domain models to help planners solve planning problems more efficiently. A crucial verification task is is to ensure that the addition of macros does not alter the functionality of the original domains. In other words, check that the modified and original planning domain models are functionally equivalent.

Planning domain model validation can also be used to check if two planning models, that allegedly represent the same domain, are actually equivalent [5]. This validation task compares two independently developed domains for the same model in critical applications and exposes any discrepancies.

Validating the equivalence of planning domain models is a challenging task. Two planning domain models could be equivalent and yet have different appearances. They could have a different number of operators. Moreover, operators and predicates could have different names between the two domains. The following definitions and theorem provides the theoretical base for our method.

Definition 3.1. Two planning domains D_1 and D_2 are **functionally equivalent** if and only if there is a bijective mapping from the predicates of D_1 to the predicates of D_2 such as when the predicates of D_1 are substituted with the predicates from D_2 as per the predicates mapping, the reach set of each domain is contained in the reach set of the other domain for any set of objects.

Definition 3.2. The **reach set** of a planning domain model D over a set of objects Obj is defined using the set of actions A as the union of all successor states of all states of the state space S that is spawned from the sets of the domain predicates and the world objects Obj. The successor states of a state $s \in S$ are produced by applying all applicable actions from A in the state s.

Theorem 3.1. *For a set of objects and two planning domain models, D_1 and D_2, with the same set of predicates, the reach set of the planning domain model D_1 is a subset of the reach set of the planning domain model D_2 if and only if the reach set of each operator from the domain D_1 is a subset of the reach set of a sequence of operators from the domain D_2.*

The condition of the same predicates can be generalised to two different sets of predicates with a proper mapping function. We omit the proof due to lack of space.

Ultimately, we intend to check whether two planning domains are functionally equivalent. First, a planner is used to find all potential macros from one domain that cover the functionality of each one of the operators in the other domain. This is done for both domains. Then, we use a SAT solver to find a consistent mapping between the predicates of the two domains while respecting the operators functionality coverage as decided in the first step. If such mapping is found, then the two domains are functionally equivalent.

The main challenge for this method is that when there is no consistent mapping for the predicates, the time and space requirement for proving that the planning domain models are not equivalent will grow exponentially with the size of the domains. However, we expect the worst-case scenarios to be very rare in real life applications, as there are necessary but not sufficient conditions that could easily disprove planning domain models equivalence for most non-equivalent domains.

4 CONCLUSION AND FUTURE WORK

In this abstract, we have summarised our goal-constrained planning domain model verification method of safety properties. An approach that increases the robustness of planning domain model verification by eliminating false positive counterexamples. We have also introduced the planning domain model validation problem and have proposed a method to check the functional equivalence of planning domains. Currently, our method can just confirm whether two domains are equivalent or not. As future work we aim to improve our algorithm to decide on the similarity between planning domain models by means of some distance measures. As a new research direction, we will investigate how to automatically debug failed test cases in test-based verification of planning-based autonomous systems to pinpoint the root cause of such failures.

ACKNOWLEDGEMENTS

This work is supported by EPSRC grant EP/P510427/1 in collaboration with Schlumberger. Section (2) summarises the work in "Anas Shrinah and Kerstin Eder, Goal-constrained planning domain model verification of safety properties, in STAIRS, (2020)". Special thanks to my supervisors Professor Kerstin Eder and Professor Derk Long for their support and guidance.

REFERENCES

[1] Diego Aineto, Sergio Jiménez, and Eva Onaindia, 'Learning strips action models with classical planning', in *Twenty-Eighth International Conference on Automated Planning and Scheduling*, (2018).

[2] Lukáš Chrpa, Mauro Vallati, and Thomas Leo McCluskey, 'On the online generation of effective macro-operators', in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, (2015).

[3] Rolf Drechsler, *ed. Advanced formal verification*, volume 122, Springer, 2004.

[4] Alfonso Gerevini and Derek Long, 'Preferences and soft constraints in PDDL3', in *ICAPS workshop on planning with preferences and soft constraints*, pp. 46–53, (2006).

[5] Thomas L McCluskey, Tiago Vaquero, and Mauro Vallati, 'Issues in planning domain model engineering', (2016).

[6] Muhammad Abdul Hakim Newton, John Levine, Maria Fox, and Derek Long, 'Learning macro-actions for arbitrary planners and domains.', in *ICAPS*, volume 2007, pp. 256–263, (2007).

[7] Hankz Hankui Zhuo and Subbarao Kambhampati, 'Action-model acquisition from noisy plan traces', in *Twenty-Third International Joint Conference on Artificial Intelligence*, (2013).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 31–32, Santiago de Compostela, Spain, August 29-30, 2020.

Improving Visual Reasoning in Visual Question Answering Systems

Kervadec Corentin[1]

Abstract. Visual Question Answering [3] (VQA), which requires answering a textual question on an image, has become an interesting test-bed for the evaluation of the reasoning and generalization capabilities of trained computational models. It deals with open questions and large varieties, and solving an instance can involve visual recognition, logic, arithmetic, spatial reasoning, intuitive physics, causality and multi-hop reasoning. In this extended abstract, we present our approaches to improve reasoning capabilities in VQA systems.

1 CONTEXT: VISUAL QUESTION ANSWERING (VQA)

High-capacity deep neural networks trained on large amount of data currently dominate methods addressing problems involving vision or language such as Visual Question Answering (VQA). The emergence of such models were possible thanks to the construction of large databases such as VQAv2 [5] (256K images with at least 3 questions for each) or GQA [7] (1.7M question-answer pairs and 180K images), both based on real images. Early work in vision and language understanding focused on separate models for each modality followed by multi-modal fusion [4, 13]. More recently, high-performing models adopt self-attention architectures and BERT-like training objectives, which complement the main task-related loss with other auxiliary losses correlated to the task [12].

The Generalisation Curse However, despite out-standing improvement made on standard VQA benchmarks, current models show a spurious tendency to over rely on harmful language and context biases present in the training corpus [2]. For instance, some models reproduce gender biases [6] observed in the training set, or others are myopic and only read half of the question [1]. One could say that the resulting VQA models "guess" the answer instead of performing a "reasoning". One the other hand, a blind model (which have only access to the question) achieves a surprisingly high accuracy – 44% of correct predictions on the VQAv2 [5] dataset. Thereby, it seems that the "generalization curse" affects as much VQA methods than the benchmarks used to evaluate them.

In that context, our work contributes to the emergence of new VQA models which are less prone to learning harmful biases and thus more reliable in real-world scenarios. Our approach is two-fold:

- Defining extra supervision signals in order to guide the model's reasoning,
- Improving the way we measure the reasoning behavior: analysing dependency toward context biases and generalisation capacities.

[1] Université de Lyon, INSA-Lyon, LIRIS UMR CNRS 5205, France; Orange Labs, Cesson-Sévigné, France; corentin.kervadec@orange.com

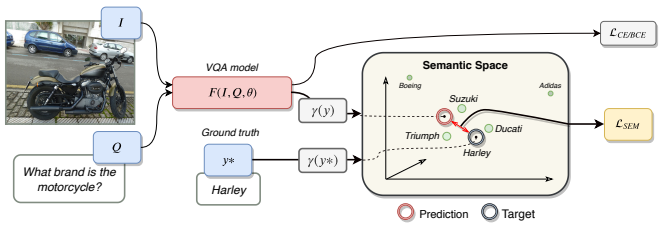

Figure 1. The model is trained to minimize the distance between the prediction and the target in a semantic space using our semantic loss $\mathcal{L}_{SEM}$.

2 CONSTRAINING VQA VIA SUPERVISION

Great efforts have been made to design efficient multimodal neural architectures [4, 13], providing helpful inductive biases for VQA, aiming to model relations between words and objects. In our work, we show the benefit of adding supplementary training objectives in order to better take advantage of these architectures. In particular we propose three approaches: a fine-grained object-word alignment, a language-grounded object detection module, and a semantic loss.

Weakly Supervised Object-Word Alignment [8] The VQA task requires different levels of understanding, from a global image-text comparison to fine-grained object-word matchings. In our paper *Weak Supervision helps Emergence of Word-Object Alignment and improves Vision-Language Tasks* [8], we show that adding an explicit fine-grained object-word alignment supervision improves the visual reasoning ability. In particular, we demonstrate the benefit of our method on VQA (GQA [7]) and Language-driven Comparison of Images (NLVR2 [11]). The latter is interesting as the NLRV2 [11] dataset has been conceived to be robust towards visual biases. Hence we demonstrate how the proposed fine-grained object-word alignment supervision really helps improving the visual reasoning. In Figure 2 we provide a comparison of the visualization of hidden attention maps learned by the model with and without the supervision.

Language-grounded object detection for VQA Current VQA models rely on an object detector to extract object-level features from the image. However, its role is limited to providing boxes and feature embeddings for general object categories, independent of the question to be answered. We then propose to add a question-grounding

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 33–34, Santiago de Compostela, Spain, August 29-30, 2020.

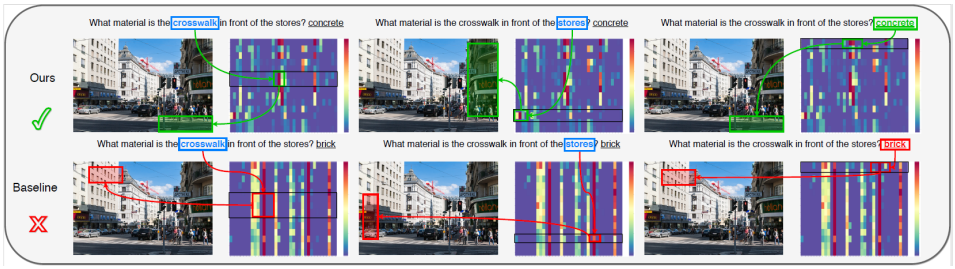

Figure 2. Visualization of the attention maps of the penultimate (=4th) inter-modality transformer. Word-object alignment does not emerge naturally for the baseline (without object-word alignment supervision), whereas our model with the proposed weakly-supervised objective learns to pay strong cross-attention on co-occurring combinations of words and objects in the scene. (Rows represent words and columns represent visual objects)

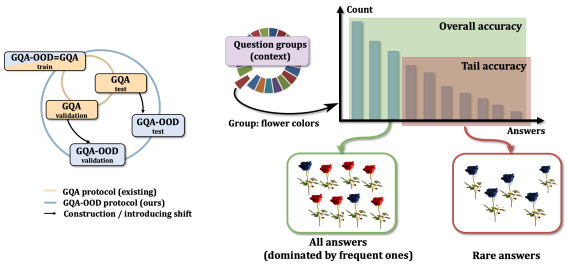

Figure 3. We re-organize the GQA dataset [7] in a fine-grained way: (a) a distribution shift allows to validate *and* evaluate in OOD settings; (b) our distribution shifts are tailored to different imbalanced question groups.

module to the standard object detector, supervised to select visual regions which are required for predicting the answer[2]. This has two main advantages: (1) filtering out useless objects which can mislead the VQA model; and (2), retrieving additional objects, which would have been missed with standard process. We test and assess the efficiency of this approach on the GQA dataset [7].

Estimating Semantic Structure for the VQA Answer Space [9] Following the same approach, *i.e.* constraining models via additional supervision signals, we propose a *Semantic Loss* [9] aiming at structuring the answer space of VQA models (cf. Figure 1). It complements the traditional classification loss by helping the model to learn semantic relations between answers. We demonstrate the benefits of the semantic loss by evaluating it on the VQA-CP [2] benchmark, conceived to measure the dependency over language biases.

3 MEASURING REASONING BEHAVIOR

During our experiments, we experience difficulties to properly measure the impact of contextual biases on the VQA performance. In particular, we observe that standard benchmarks [5, 7] repeat in the test set the harmful biases observed in the training, leading to misleading scores and covering the true behavior of VQA models. Then, in our paper *Roses are Red, Violets are Blue... But Should VQA expect Them To?* [10], we propose a large-scale study revealing that several state-of-the-art VQA models fails to address questions involving infrequent concepts. From this result, we design a new VQA benchmark named GQA-OOD based on a fine-grained re-organization of the GQA [7] dataset and dedicated to the evaluation of the reasoning and generalisation behavior (cf. Figure 3) [3]. Our study leaves a big room for improvement for bias-reduction methods and we hope it will contribute to the emergence of new VQA models.

Future work We aim to pursue the analysis of reasoning behavior in VQA. Indeed, one of the major drawback of current studies is the lack of diagnosis explaining why and how VQA systems works. Better understanding VQA models will push forward the emergence of new models capable of what we call "visual reasoning".

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh, 'Analyzing the behavior of visual question answering models', in *EMNLP*, (2016).
[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi, 'Don't just assume; look and answer: Overcoming priors for visual question answering', in *CVPR*, (2018).
[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, 'Vqa: Visual question answering', in *ICCV*, (2015).
[4] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord, 'Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection', in *AAAI*, (2019).
[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, 'Making the v in vqa matter: Elevating the role of image understanding in visual question answering', in *CVPR*, (2017).
[6] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach, 'Women also snowboard: Overcoming bias in captioning models', in *ECCV*, (2018).
[7] Drew A Hudson and Christopher D Manning, 'Gqa: A new dataset for real-world visual reasoning and compositional question answering', in *CVPR*, (2019).
[8] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf, 'Weak supervision helps emergence of word-object alignment and improves vision-language tasks', *ECAI*, (2019).
[9] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf, 'Estimating semantic structure for the vqa answer space', *arXiv*, (2020).
[10] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf, 'Roses are red, violets are blue... but should vqa expect them to?', *arXiv*, (2020).
[11] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi, 'A corpus of natural language for visual reasoning', in *ACL*, (2017).
[12] Hao Tan and Mohit Bansal, 'Lxmert: Learning cross-modality encoder representations from transformers', *arXiv*, (2019).
[13] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian, 'Deep modular co-attention networks for visual question answering', in *CVPR*, (2019).

[2] Submitted to an international conference (still under review).

[3] Available at `https://github.com/gqaood/GQA-OOD`

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 33–34, Santiago de Compostela, Spain, August 29-30, 2020.

Stylistic Dialogue Generation in Narratives

Weilai Xu[1]

Abstract. In this report, this PhD programme is presented through a brief literature review. The main goal of the PhD programme is aiming to explore the influence of generating dialogues with narrative knowledge for presenting style variety using deep learning techniques. The progress to date shows that our approach is able to provide style variety for generated dialogues from the perspective of narratives based on personality features.

1 INTRODUCTION

Dialogue is an important way to convey and exchange information in daily lives but also translates to virtual story world. A story or a narrative is an ordered sequence of connected events in which one or more characters (or entities) participate. Normally, a narrative can be divided into two constituent parts: plot (or story) and discourse. A plot is made up of several elements composing the resulting narrative, such as locations, characters, their relationships, narrative actions, etc. Finally discourse denotes an instantiation of the plot through a means of presentation which can be expressed in different ways (e.g. visual, text, sound, etc.). In narratives, including a lot of character interactions, dialogue usually plays an essential role to express the plot and the personalities of characters.

Dialogue generation is a problem of renewed interest in AI, in which many contributions have been made recently due to the rise of deep learning techniques [6, 8, 9] . However, most of these works deliver content information through the generated sentences, focusing on the aspect of "what" to say, rather than the aspect of "how" to say it, which relates to combining stylistic information with the semantic content. In this case, we argue that there is still a need for work to be done to meet this promising target.

By incorporating additional features, some works have generated utterances and sentences in natural language using deep learning, providing style variety to some extents, such as speaker profile [6, 8] , personality [5, 8] , and sentiment or emotion [3, 4]. Although such features can offer positive outcomes for the generation of varying styles in more generic dialogues, it remains difficult to apply these features into a successful approach for stylistic narrative dialogue generation. Primarily, these works alter the conditions of a single utterance or sentence during generation, which means these features affect word-level local expression within individual sentences. However, These features are mostly explored on the perspective of interlocutors, which cannot reflect the authoring actions and designs by authors in storytelling. Additionally, these considered features are not narrative specific, but instead are simply extracted from and applied within task-oriented domains (e.g. customer service and restaurant booking). Only very few of these features would be relevant for narrative domains, such as TV series, films, games, etc.

[1] Bournemouth University, UK, email: wxu@bournemouth.ac.uk

Therefore, based on the updated research, the goal of this PhD programme is to explore the potential of how to generate dialogues within the scope of narratives with various styles. In order to reach this goal, some research questions are proposed as follows:

1. **Why is it promising that incorporating narrative-based features for generating stylistic dialogue?**
2. **What approach can be used for representing and adding narrative-based features into the neural dialogue system in order to provide various styles?**
3. **What differences do the additional features make to the generated dialogues using this approach, and how to evaluate them?**

Since it is an interdisciplinary PhD programme, the academic works on storytelling, narratives, natural language generating and machine learning have been widely reviewed and researched, for discovering a combination point where the dialogue generation can be boosted by the knowledge in narratives. Therefore, the three main contributions are supposed to be made at the end of the PhD are listed as follows,

1. A review discussing the reasons that neural dialogue generation approach could be and should be leveraged in narrative area.
2. An approach that incorporate narrative-based features (e.g. personality) into deep neural dialogue system and an experiment for evaluating this approach.
3. A dataset which contains great amount dialogues along with narrative features on different aspect, parsed from IMSDb.

2 PROGRESS TO DATE

So far a mid-term major report was completed with two research works published amid the PhD programme [10, 11]. In these two publications, we have discussed and analysed the existing related theories and approaches in both computational narrative and natural language generation communities, aiming for discovering and exploring the potential connections between both. During this time, we found that most dialogue generation research works require further knowledge of narratives, which is the main motivation of this PhD in order to improve the approach for more stylistic dialogue generation within the context of narratives.

In [11], we introduce a stylistic dialogue modelling based on different linguistic features on the aspect of genre diversity and the relationship between characters. We intend to incorporate these narrative-based feature model into neural dialogue generation process. Also, we design a workflow using LSTM neural model and explain how the features are fed into the system theoretically. We build the foundation in this paper for the further implementation stage.

In [10], we explore the influence of different personality-based features for dialogue generation. Character is a essential element in

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 35–36, Santiago de Compostela, Spain, August 29-30, 2020.

storytelling that responsible for progressing the storyline and conveying the intention of authors. Particularly, personality represents the figure that authors create, reflecting the authorisation to the biggest extent according narrative theory [7]. Therefore, we implement LSTM-based neural dialogue system, and trained the system on the self-processed corpus with dialogue text and personality labels, which are collected from the screenplays in IMSDb. The results from evaluation show that our approach has the capacity to provide style variety of generated utterances across different personality trait combinations by the metrics of edit distance and cosine similarity. The generated utterances can be altered according to different certain personality trait combination using trained dialogue generation system on narrative perspective (partial results shown as Figure 1). For example, we observe that utterances with extravert personality (i.e. *high* extraversion) tend to have more variety than with introvert personality (i.e. *low* extraversion).

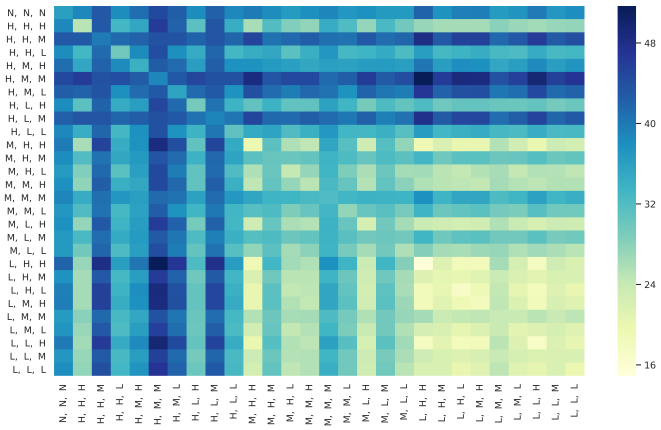

Figure 1. Mean edit distance between two personality trait combinations pair-wisely of 30 test utterances in the Drama genre. Each row and column labels contain the labels of three personality traits (extraversion, emotional stability, and agreeableness). These labels are ordered by the label *High*, *Medium*, and *Low* for each trait. A blank control combination without trait label is placed on the first row and column, labelled *None*.

The approach has also been implemented using transformer architecture and pre-trained GPT-2 language model. A well-designed user study has been launched for evaluating to what extent do the additional features affect the quality of generated dialogues from the perspective of narratives. In the user study, each dialogue is reviewed on three dimensions for evaluating dialogue quality (Grammar, Consistency, and Readability), and the personality setting is identified for the evaluation of approach reliability. By analysing the results, it is promising that the proposed questions could be answered properly.

It is a novel experimental approach by which we explore the potential of incorporating narrative-based features into neural dialogue generation system using deep learning and big data, differing from most previous research in narrative community using plan-based and rule-based methodology (e.g. [1, 2]). Therefore, the main contributions of this PhD are applying different narrative-based features in the neural dialogue generation procedure, as well as to evaluate and find which feature(s) provide more influence to the generated dialogues and provide more styles in the context of narratives.

However, based on progress so far, we noticed challenges in the research. First, it is difficult to define and quantify narrative-based features, while most of the existing corpora in natural language contain the features which are easy to classify in taxonomy, such as speakers' gender and age, the topics of the dialogue, etc. Therefore, we have to define our own narrative-based features derived from narrative theories and create corpus with such features. Second, we found that there is a lack of effective existing methods for representing these features as the formations for the neural networks. To address this limitation, we are working on adapting existing methods to represent features with the utterances and embed them into neural dialogue system.

3 DISCUSSION

In this report, we introduce an overview of the whole PhD programme with three research questions, after discussing and analysing the research gaps from current related work. We also briefly present and explain our up-to-date progress to answer those questions.

It is widely witnessed that neural language model has achieved huge improvements in natural language process and generation and text is also a important way for presenting narratives. Therefore, we believe it is possible to combine them to generate narrative-based stylistic dialogue with the support of advanced techniques.

We believe that the observations and discoveries from this PhD programme could be a start and a tryout to apply deep learning technique and big data to boost narrative dialogue generation. And we also believe that our research can be applied in plenty of potential scenarios, such as helping the authors creating huge amount of conversations between different characters by popping utterance options corresponding the character settings or previous storyline progress.

ACKNOWLEDGEMENTS

The author also has a paper accepted in the main ECAI2020 conference [10].

REFERENCES

[1] Kevin K Bowden, Grace I Lin, Lena I Reed, Jean E Fox Tree, and Marilyn A Walker, 'M2d: Monolog to dialog generation for conversational story telling', in *International Conference on Interactive Digital Storytelling*, pp. 12–24. Springer, (2016).

[2] Marc Cavazza and Fred Charles, 'Dialogue generation in character-based interactive storytelling', in *Proceedings of the First AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pp. 21–26. AAAI Press, (2005).

[3] Jessica Ficler and Yoav Goldberg, 'Controlling linguistic style aspects in neural language generation', in *Proceedings of the Workshop on Stylistic Variation*, pp. 94–104, (2017).

[4] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, 'Affect-lm: A neural language model for customizable affective text generation', *arXiv preprint arXiv:1704.06851*, (2017).

[5] Jonathan Herzig, Michal Shmueli-Scheuer, Tommy Sandbank, and David Konopnicki, 'Neural response generation for customer service based on personality traits', in *Proceedings of the 10th International Conference on Natural Language Generation*, pp. 252–256, (2017).

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

[7] Robert McKee, 'Substance, structure, style, and the principles of screenwriting', (1997).

[8] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu, 'Assigning personality/profile to a chatting machine for coherent conversation generation.', in *IJCAI-ECAI*, pp. 4279–4285, (2018).

[9] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, 'Sequence to sequence learning with neural networks', in *Advances in neural information processing systems*, pp. 3104–3112, (2014).

[10] Weilai Xu, Fred Charles, Charlie Hargood, Feng Tian, and Wen Tang, 'Influence of personality-based features for dialogue generation in computational narratives', in *ECAI*, (2020).

[11] Weilai Xu, Charlie Hargood, Wen Tang, and Fred Charles, 'Towards generating stylistic dialogues for narratives using data-driven approaches', in *International Conference on Interactive Digital Storytelling*, pp. 462–472. Springer, (2018).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 35–36, Santiago de Compostela, Spain, August 29-30, 2020.

Improving Robustness of Emotion Recognition in Multi-Accent and Multilingual Speech

Aaron Keesing[1]

Abstract. Speech emotion recognition (SER) is receiving increasing interest due to the popularity of speech-enabled devices. The aim of SER is to estimate the emotional state of a person from the acoustic and linguistic content of their speech. While much work has been done to improve SER on monolingual native speech, less work has been done in the multilingual case, and cases where the speakers are L2 speakers. This research aims to improve the robustness of SER in multi-accent and multilingual scenarios. For the multi-accent case we are particularly interested in including accents of non-native speakers that influence their prosodic patterns and perceived emotion. To solve this issue we intend to use three machine learning techniques: automatic data augmentation, self-training, and transfer learning.

1 INTRODUCTION

Speech emotion recognition (SER) is a sub-field of artificial intelligence and affective computing that aims to predict the emotional state of a person from their speech. It is cross-disciplinary by nature, at the intersection of human-computer interaction (HCI), AI, speech processing, psychology, affective computing, and having a wide range of applications.

A definition of emotion proposed by Cowie et al. [3], defines emotion in the negative as, "whatever is present in most of life, but absent when people are emotionless," noting that 'emotionless' tends to have a stricter definition than 'emotional'. They also discuss the issue of defining 'realistic' or 'naturalistic' emotions. Realistic emotions are not prompted and tend to better represent the occurrence of emotional expression in everyday life. On the other hand, acted emotion may be more contrived.

Typically, a discrete list of basic emotions is used, such as Ekman's [4] 'big six'. Continuous emotional models are most commonly implemented as a two-dimensional *arousal-valence* space; arousal measures the degree of activity that an emotion excites, while valence measures how positive or negative the overall feeling is. Additional dimensions have also been used in emotion recognition, such as dominance and power [2].

An important issue to consider for multilingual research is any cultural variation in emotional perception and expression. The research indicates that there are both similarities and differences in emotional expression and perception [12] across cultures. People in disparate cultures are able to correctly identify the emotions of a people in another culture based on facial images [5], or speech [11], however their accuracy is greater within-culture.

There are a several datasets that have been used for emotion recognition, but they have two main problems. Firstly, the datasets are quite sparse in the sense that the space of possible emotional expressions is quite vast, while each dataset typically focuses on a specific language and a few fixed emotions. The second issue is the compatibility of datasets for comparison or combination; different datasets may use different emotion words for discrete labelling, or may differ in the type of annotation. There are additional issues when combining corpora for testing multilingual SER because the meaning of emotion words may be different across languages. An issue relevant to accent-robustness is that databases usually contain speech only of native speakers in a similar geographic location, thus having similar accents. The converse is that a reasonable amount of accented speech data is available, but without emotional annotations. One part of our research is to gather emotional annotations of different accented speech using crowd-sourcing.

Features used in SER can be linguistic but are more commonly acoustic, and can be the raw waveform itself, the spectrogram, or specific descriptive features. Descriptive features start with low-level descriptors (LLDs) which represent different aspects of the speech signal (e.g. energy, pitch, formants, MFCCs, etc.) and are calculated in short duration (≈25ms) frames from the signal. From LLDs, functionals (such as mean, max, standard deviation, regression coefficients) are calculated across frames, segments or utterances. Sequence models may use the LLDs directly, while classifiers use a vector of functionals calculated. Four standard feature sets in SER are IS09 [14], IS13 [13], GeMAPS, and eGeMAPS [6]. More recently, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can be trained in an end-to-end fashion either directly from the time-domain signal, or from the spectrogram. Alternatively, unsupervised learning can be used to automatically learn representative features that encode the acoustic signal, which can subsequently be used in another classifier.

2 RESEARCH GOALS

This research has two main broad research goals:

- Create emotion recognition models that are robust against changes in speech accent, including from non-native speakers.
- Create emotion recognition models that are robust against changes in language and culture.

In particular, for the first research goal, the aim is to correctly predict the emotional state of speakers with different accents of the same language. For the second research goal, the aim is accurate emotion recognition in multiple languages. Robustness means that a model does not perform significantly worse on novel accents/languages that it would when training on those accents/languages. In both cases the criteria for success are achieving accuracy on unseen test data that

[1] University of Auckland, New Zealand, email: akee511@aucklanduni.ac.nz

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 37–38, Santiago de Compostela, Spain, August 29-30, 2020.

is comparable to what would be achieved in the within-corpus case, which should be at least that of standard baseline models such as SVM. It will be more difficult to compare to 'state of the art' multi-corpus results, unless those results use the same datasets, emotional annotations, and evaluation procedures, which is less likely given the greater variability when using multiple corpora.

3 METHODOLOGY

Thus far, some baseline within-corpus results with various feature sets and model types have been tested, serving as baseline for future research and comparison with multi-corpus testing. Development tools and an emotion recognition testing framework have been developed and made open source[2].

In this research, speech data in multiple accents, including L2 speakers, will be gathered from pre-existing datasets, and crowd-sourcing will be used to provide emotional annotations. This will supplement existing datasets of emotional speech in different accents. The crowd-sourcing will be done through a crowd-sourcing platform, where emotional annotation tasks can be created and completed by volunteers for a small payment. Once a task is completed, it will be determined as *satisfactory* or *unsatisfactory* based on both self-consistency and a minimum level of agreement with other raters [1]. For greater flexibility in testing and for compatibility when combining with other emotion datasets, annotations will include both emotion words and the three emotion dimensions of valence, arousal and dominance. We will follow a similar procedure to [9], in which reference clips are inserted periodically in order to determine whether an annotator's performance is decreasing, and conclude the annotation session if performance drops below a certain threshold. Self-assessment manikins [7] will be used to aid in the dimensional annotation, where annotators can select pictorial representation of the emotional dimensions on a five-point scale. Once multiple annotations are collected for each clip, a majority vote will be used on the discrete labels as the 'gold standard' for training a classifier, and the mean dimensional values after annotator normalisation will be used for regression.

When training models, data augmentation such as SpecAugment [10] will be used to slightly non-linearly warp the the spectrogram along the time axis. Another data augmentation technique is vocal-tract length perterbation (VTLP) [8] which randomly scales the frequency axis linearly. Care must be taken, however, that the spectrogram is not modified so much as to increase the variance in prosodic features between emotions. Masking in time and frequency, and adding white noise, can improve robustness against differing recording conditions and audio quality.

The main focus will be on discrete emotion classification. Initially, for better comparability to previous results, emotion classes or arousal/valence will be mapped to binary arousal and valence. When combining corpora in the same language, a common set of emotional labels will be used if possible.

After testing using only data augmentation techniques, self-trained models will be tested with data augmentation. This is a *weakly*-supervised technique whereby a small amount of labelled data is used to train a model, after which it generates labels for unlabelled data which it uses to train itself, and so on, in an iterative fashion, in order to bootstrap the model on mostly unlabelled data. This is ideal for emotion recognition due to the sparsity of annotated emotional speech data but relative prevalence of general speech data.

[2] `https://github.com/agkphysics/emotion`

We also aim to test the improvement that transfer learning gives over the methods described above, to see to what extent transfer learning is required for robust emotion recognition across accents and languages. Transfer learning uses a small amount of labelled or unlabelled target data to change the model's distribution towards the target distribution.

ACKNOWLEDGEMENTS

This project is being completed under the supervision of Assoc. Prof. Ian Watson and Prof. Michael Witbrock.

REFERENCES

[1] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, 'MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception', *IEEE Transactions on Affective Computing*, **8**(1), 67–80, (January 2017).

[2] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder, ''FEELTRACE': An instrument for recording perceived emotion in real time', in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, (2000).

[3] Roddy Cowie, Naomi Sussman, and Aaron Ben-Ze'ev, 'Emotion: Concepts and Definitions', in *Emotion-Oriented Systems: The Humaine Handbook*, eds., Roddy Cowie, Catherine Pelachaud, and Paolo Petta, Cognitive Technologies, 9–30, Springer, Berlin, Heidelberg, (2011).

[4] Paul Ekman, 'An argument for basic emotions', *Cognition and Emotion*, **6**(3-4), 169–200, (May 1992).

[5] Paul Ekman and Wallace V. Friesen, 'Constants across cultures in the face and emotion', *Journal of Personality*, **17**(2), 124–129, (February 1971).

[6] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, 'The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing', *IEEE Transactions on Affective Computing*, **7**(2), 190–202, (April 2016).

[7] M. Grimm and K. Kroschel, 'Evaluation of natural emotions using self assessment manikins', in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pp. 381–385, (November 2005).

[8] Navdeep Jaitly and Geoffrey E. Hinton, 'Vocal tract length perturbation (VTLP) improves speech recognition', in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, volume 117, (2013).

[9] R. Lotfian and C. Busso, 'Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings', *IEEE Transactions on Affective Computing*, 1–1, (2017).

[10] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, 'SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition', in *Interspeech 2019*, pp. 2613–2617. ISCA, (September 2019).

[11] Klaus R. Scherer, Rainer Banse, and Harald G. Wallbott, 'Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures', *Journal of Cross-Cultural Psychology*, **32**(1), 76–92, (January 2001).

[12] Klaus R. Scherer and Harald G. Wallbott, 'Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning.', *Journal of Personality*, **66**(2), 310–328, (February 1994).

[13] Björn Schuller and Anton Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*, John Wiley & Sons, 2013.

[14] Björn Schuller, Stefan Steidl, and Anton Batliner, 'The INTERSPEECH 2009 emotion challenge', in *10th Annual Conference of the International Speech Communication Association*, pp. 312–315, Brighton, United Kingdom, (September 2009).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 37–38, Santiago de Compostela, Spain, August 29-30, 2020.

Fuzzy Quantified Protoforms for Data-To-Text Systems: a new model with applications

Andrea Cascallar-Fuentes[1]

Abstract. A huge quantity of data is generated every second which contains valuable information we need to extract. In this context, the tools that allow to communicate the results of these analyses in a comprehensible way to human decision makers are not so developed. This is where data-to-text (D2T) systems, a discipline that focuses on the automatic generation of texts from various sources of numerical or symbolic data, is presented as a useful technology. D2T systems are capable of processing large amounts of numerical data, generating comprehensible texts that contain relevant information for users.

The objective of this thesis is threefold: i) the proposal of new models composed by fuzzy protoforms that include geographical information. ii) An empirical assessment and comparison of the impact of the selection the fuzzy quantification method to assess their empirical behavior when applied for the evaluation of fuzzy quantified sentences and iii) the design of a model to manage the imprecision of temporal knowledge.

1 INTRODUCTION

Organizations generate and consume large amounts of data that have real value insofar as they can be transformed into relevant and useful information, which can be effectively transferred so that it is considered in the decision-making processes. In this context, although the tools for data analysis are already very common, the tools that allow to communicate the results of these analysis in a comprehensible way to human decision makers are not so developed. This is where data-to-text (D2T) systems, a discipline that focuses on the automatic generation of texts from various sources of numerical or symbolic data, is presented as an emerging technology of undoubted usefulness. Within the more general area of the Natural Language Generation, D2T systems [16] are capable of processing large amounts of numerical data, converting them into texts that contain relevant and understandable information for users, so that information can be automatically extracted from said data and communicated in an intuitive way. An important task within these systems is the analysis of data made to obtain the basic pieces of information that will be later integrated into the texts. This phase is called data processing or content determination [22] and is one of the objectives of this thesis.

At the same time, in the fuzzy logic and soft computing fields many approaches were proposed to generate describe data using linguistic terms [28], for instance, Linguistic Descriptions of Data (LDD) [18] which summarize in a linguistic form one or more numerical variables and their values, using the general notion of protoform [27]. Protoforms can follow several structures being unary and binary linguistic descriptions the most common in the literature.

Unary descriptions have the following structure: *"Q X are A"*, where Q is a linguistic quantifier, X is a linguistic variable defined on a given referential and A is a fuzzy linguistic value (property) of X. For instance, in *"Most temperatures are normal," "Most"* is the quantifier, *"temperatures"* is the linguistic variable and *"normal"* is a linguistic value of *temperatures*.

Binary descriptions follow the structure *"Q DX are A"* where an additional fuzzy property *"D"* is defined on the same referential of X. For example, in *Most temperatures in the North are normal "North"* is the additional fuzzy property (D).

Evaluating a quantified sentence involves obtaining its truth value on a given data set. It is obtained by calculating the compatibility between the number of elements in the referential which fulfill the sentence (its cardinality) and the quantifier in the sentence. Therefore, this measure depends on the data, the quantifier definition, the linguistic terms defined from the properties in the referential and the quantification model used for evaluating the sentence. There exist several quantification methods in the literature ([3, 25, 26] among others) that differ from each other in the way they calculate the truth value.

2 OBJECTIVES AND RESULTS

The objective of this thesis is threefold: i) on one side, we aim to improve and extend content determination in D2T systems by introducing new techniques based on artificial intelligence for the representation of imprecise knowledge and intelligent search. Specifically, we will propose new models composed by fuzzy protoforms that include information and geographical relationships. We will design scalable algorithms that allow their extraction on several types of data. To perform this task, in the literature both heuristic [5, 21] and meta-heuristic [6, 9] approaches have been proposed. We will consider, on the one hand, meta-heuristic approaches to obtain a good compromise between the quality of the solution and the computational cost of the solution. On the other, we will also design distributed versions of them that can be applied to large data-sets. In [15] we presented the generation of fuzzy quantified statements with a purpose based on the Simulated Annealing algorithm.

ii) Our second objective is to measure and compare the impact of the selection of the fuzzy quantification method to assess their empirical behavior when applied for the evaluation of fuzzy quantified sentences. Several fuzzy quantification models have been proposed in the literature and were later studied from a theoretical perspective in terms of the properties they fulfill [1, 3, 7, 8, 10, 11, 12, 13, 17, 24]. An extensive list of properties has been described considering different aspects that help to characterize the behavior of the fuzzy quan-

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain, email: andrea.cascallar.fuentes@usc.es

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 39–40, Santiago de Compostela, Spain, August 29-30, 2020.

tification models. From this perspective, all the fuzzy quantification models exhibit different behavior, since all of them fulfill different properties. But an empirical comparison between their behavior has not been performed yet, therefore our aim is to experimentally test whether there are significant differences between the most widely used unary and binary fuzzy quantification models used in quantified sentences. Results of these experiments are presented in [4, 19].

Finally, our last objective is iii) the design of a model which cover the data-to-text pipeline [22] of fuzzy quantified statements to describe temporal series. Due to the importance of the temporal dimension of the data, we will define a temporal ontology with the temporal concepts definitions to process temporal knowledge. In [14] an approach to manage the imprecision of temporal knowledge is proposed. Also some approaches proposed the definition of languages and ontologies to manage temporal data [2, 20, 23].

This proposed model will be validated in two real application areas: meteorology and cardiac rehabilitation.

3 FUTURE WORK

As future work, related to our first objective, our aim is *i)* to consider other meta-heuristic algorithms, *ii)* consider other types of protoforms and *ii)* develop the entire data-to-text pipeline in order to create a fluent text. Regarding to our second objective, a possible future work is *i)* the consideration of more fuzzy quantification methods and *ii)* including new protoform types. Finally, regarding to our third objective, the future work is its development.

ACKNOWLEDGEMENTS

I want to recognize the role and support of my PhD advisors, Alberto Bugarín-Diz and Alejandro Ramos-Soto.

We have an accepted paper in the main ECAI2020 conference [4].

REFERENCES

[1] Senén Barro, Alberto Bugarín, Purificación Cariñena, and Félix Díaz-Hermida, 'A framework for fuzzy quantification models analysis', *IEEE Trans. Fuzzy Systems*, **11**(1), 89–99, (2003).

[2] Senén Barro, Roque Marín, José Mira, and Alfonso R Patón, 'A model and a language for the fuzzy representation and handling of time', *Fuzzy sets and Systems*, **61**(2), 153–175, (1994).

[3] Miguel Delgado Calvo-Flores, Daniel Sánchez, and M. Amparo Vila, 'Un método para la evaluación de sentencias con cuantificadores lingüísticos', in *Actas del VIII Congreso Español sobre Tecnologías y Lógica Fuzzy: Pamplona, 8-10 de septiembre de 1998*, pp. 193–198. Departamento de Automática y Computación, (1998).

[4] Andrea Cascallar-Fuentes, Alejandro Ramos-Soto, and Alberto Bugarín-Diz, 'An experimental study on the use of fuzzy quantification models for linguistic descriptions of data', in *24th European Conference on Artificial Intelligence*, (2020).

[5] Rita Castillo-Ortega, Nicolás Marín, and Daniel Sánchez, 'A fuzzy approach to the linguistic summarization of time series', *J. Multiple Valued Log. Soft Comput.*, **17**(2-3), 157–182, (2011).

[6] Rita Castillo-Ortega, Nicolás Marín, Daniel Sánchez, and Andrea G. B. Tettamanzi, 'Linguistic summarization of time series data using genetic algorithms', in *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology, EUSFLAT 2011, Aix-Les-Bains, France, July 18-22, 2011*, eds., Sylvie Galichet, Javier Montero, and Gilles Mauris, pp. 416–423. Atlantis Press, (2011).

[7] Miguel Delgado, M. Dolores Ruiz, Daniel Sánchez, and M. Amparo Vila, 'Fuzzy quantification: a state of the art', *Fuzzy Sets and Systems*, **242**, 1–30, (2014).

[8] Miguel Delgado, Daniel Sánchez, Maria J. Martín-Bautista, and M. Amparo Vila, 'A probabilistic definition of a nonconvex fuzzy cardinality', *Fuzzy Sets and Systems*, **126**(2), 177–190, (2002).

[9] Carlos Alberto Donis Diaz, Rafael Bello, and Janusz Kacprzyk, 'Using ant colony optimization and genetic algorithms for the linguistic summarization of creep data', in *Intelligent Systems'2014 - Proceedings of the 7th International Conference Intelligent Systems IEEE IS'2014, September 24-26, 2014, Warsaw, Poland, Volume 1: Mathematical Foundations, Theory, Analyses*, eds., Plamen P. Angelov, Krassimir T. Atanassov, Lyubka Doukovska, Mincho Hadjiski, Vladimir Simov Jotsov, Janusz Kacprzyk, Nikola K. Kasabov, Sotir Sotirov, Eulalia Szmidt, and Slawomir Zadrozny, volume 322 of *Advances in Intelligent Systems and Computing*, pp. 81–92. Springer, (2014).

[10] Félix Díaz-Hermida, *Modelos de cuantificación borrosa basados en una interpretación probabilıstica y su aplicación en recuperación de información*, Ph.D. dissertation, PhD thesis, Universidade de Santiago de Compostela, 2006.

[11] Félix Díaz-Hermida, David E. Losada, Alberto Bugarín, and Senén Barro, 'A probabilistic quantifier fuzzification mechanism: The model and its evaluation for information retrieval', *IEEE Trans. Fuzzy Systems*, **13**(5), 688–700, (2005).

[12] Félix Díaz-Hermida, Marcos Matabuena, and Juan Carlos Vidal, 'The FA quantifier fuzzification mechanism: analysis of convergence and efficient implementations', *CoRR*, **abs/1902.02132**, (2019).

[13] Félix Díaz-Hermida, Martin Pereira-Fariña, Juan Carlos Vidal, and Alejandro Ramos-Soto, 'Characterizing quantifier fuzzification mechanisms: A behavioral guide for applications', *Fuzzy Sets and Systems*, **345**, 1–23, (2018).

[14] Didier Dubois and Henri Prade, 'Processing fuzzy temporal knowledge', *IEEE Trans. Syst. Man Cybern.*, **19**(4), 729–744, (1989).

[15] Andrea Cascallar Fuentes, Alejandro Ramos Soto, and Alberto José Bugarín Diz, 'Descripciones lingüísticas de datos de observación meteorológica usando temple simulado', in *XIX Congreso Español sobre Tecnologías y Lógica Fuzzy,*, pp. 469–474, (2018).

[16] Albert Gatt and Emiel Krahmer, 'Survey of the state of the art in natural language generation: Core tasks, applications and evaluation', *J. Artif. Intell. Res.*, **61**, 65–170, (2018).

[17] Ingo Glöckner, 'DFS–an axiomatic approach to fuzzy quantification', *TR97-06, Techn. Fakultät, Univ. Bielefeld*, **2**(3), 10, (1997).

[18] Janusz Kacprzyk and Ronald R Yager, 'Linguistic summaries of data using fuzzy logic', *International Journal of General System*, **30**(2), 133–154, (2001).

[19] Carlos Heble Lahera, Andrea Cascallar Fuentes, Alejandro Ramos Soto, and Alberto Bugarín Diz, 'Empirical study of fuzzy quantification models for linguistic descriptions of meteorological data', in *2020 IEEE International Conference on Fuzzy Systems*, (2020).

[20] Feng Pan and Jerry R Hobbs, 'Time in owl-s', in *Proceedings of the AAAI Spring Symposium on Semantic Web Services*, pp. 29–36, (2004).

[21] Alejandro Ramos-Soto, Alberto José Bugarín Diz, Senén Barro, and Juan Taboada, 'Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data', *IEEE Trans. Fuzzy Syst.*, **23**(1), 44–57, (2015).

[22] Ehud Reiter, 'An architecture for data-to-text systems', in *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pp. 97–104. Association for Computational Linguistics, (2007).

[23] Majid RobatJazi, Marek Reformat, and Witold Pedrycz, 'Ontology-based framework for reasoning with fuzzy temporal data', in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, SMC 2012, Seoul, Korea (South), October 14-17, 2012*, pp. 2030–2035. IEEE, (2012).

[24] M. Amparo Vila, Juan-Carlos Cubero, Juan-Miguel Medina, and Olga Pons, 'Using owa operator in flexible query processing', in *The ordered weighted averaging operators*, 258–274, Springer, (1997).

[25] Ronald R. Yager, 'On ordered weighted averaging aggregation operators in multicriteria decisionmaking', *IEEE Trans. Systems, Man, and Cybernetics*, **18**(1), 183–190, (1988).

[26] Lotfi A Zadeh, 'A computational approach to fuzzy quantifiers in natural languages', in *Computational linguistics*, 149–184, Elsevier, (1983).

[27] Lotfi A Zadeh, 'A prototype-centered approach to adding deduction capability to search engines-the concept of protoform', in *Intelligent Systems, 2002. Proceedings. 2002 First International IEEE Symposium*, volume 1, pp. 2–3. IEEE, (2002).

[28] Lotfi A Zadeh and J Kacprzyk, 'CWW in Information/Intelligent Systems, vol. 1 foundations, vol. 2. aplications', *Physica-Verlag, Heidelberg*, (1999).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 39–40, Santiago de Compostela, Spain, August 29-30, 2020.

A Two-Level Item Response Theory Model to Evaluate Automatic Speech Recognition Systems

Chaina Santos Oliveira[1]

Abstract. Automatic speech recognition (ASR) systems have become popular in different applications (e.g., in virtual assistants and intelligent interfaces). Ideally, ASR systems should be tested under various scenarios by adopting diverse speech test data to simulate real users' variability. Such data should cover, for instance, varied test sentences and speakers with different accents. Relying on speech test data recorded using human speakers is time-consuming and expensive. An alternative is to adopt text-to-speech (TTS) tools to synthesize speeches from a set of sentences given as input and virtual speakers provided by these tools. The ASR under test then receives the synthesized speeches as test data and transcribes them. Finally, the errors observed in the derived transcriptions are considered for ASR evaluation.

Despite the availability of TTS tools, not all synthesized speeches have the same quality for all speakers and sentences. So, before testing the ASR, it is essential to evaluate the speakers' usefulness and to determine which sentences are more relevant for ASR evaluation. In this context, we propose the application of Item Response Theory (IRT) to evaluate ASR systems, speech synthesizers and sentences simultaneously.

IRT is a paradigm developed in psychometrics to estimate the cognitive ability of human respondents based on their responses to items with different levels of difficulty [3, 2]. Recently, this methodology of evaluation has been used in other contexts, including in the evaluation of Artificial Intelligence (AI) systems. In supervised learning for instance, IRT has been adopted by [1], [4] and [5] to evaluate the ability of classifiers based on their answers to test instances grouped by difficulty. In [4] and [5], it is discussed the importance of analyzing particular situations in which good techniques fail (e.g., a classifier with good performance does not hit an instance that a poor classifier does). So, it is not always robust to evaluate a classifier just based on the number of instances it correctly solves. It is also important to consider the difficulty of instances classified by the models under test.

Based on the potential application of IRT in the evaluation of AI systems, in our research we adopted IRT for evaluation the speech recognition and speech synthesis context. A two-level IRT model is proposed to identify the ability of the ASR systems under test, the difficulty of sentences adopted for testing and the quality of speakers utilized for synthesis. In the first level, an item is a synthesized speech, which is produced from a given sentence and a speaker. In turn, a respondent is an ASR system. Each response is the transcription word accuracy ($WAcc$) observed when a synthesized speech is adopted for testing the ASR. An IRT model identifies latent patterns of responses to estimate each synthesized speech's difficulty and the ability of each ASR system. In the second level, the difficulty of the synthesized speech is decomposed into two latent factors: the difficulty of the sentence itself and the speaker's quality. In this sense, the difficulty of a synthesized speech is high when it is generated from a difficult sentence and a poor speaker.

So, the fundamental idea in the proposal is to model a response (transcription accuracy) by a two-level IRT model, in which: (1) the response depends directly on the ASR system ability and on the synthetized speech difficulty; (2) in turn, the synthetized speech difficulty depends on the speakers quality and on the sentences difficulty and discrimination. In this sense, a difficulty sentence and bad speaker will result on a synthetized speech which will be hard to be transcribed. Then a bad response is likely to be observed when hard audios are given to low ability ASR systems. In our work, we adopted the β^3-IRT model proposed in [1] at each level, since it is more adequate to deal with continuous responses (in this case, $WAcc$).

A case study was developed to verify the viability of the proposal. In this study, four ASR systems were adopted to transcribe synthesized speeches from 100 benchmark sentences (e.g., "there was an auto accident") using 75 speakers (from four TTS tools). IRT was applied to estimate the abilities of ASR systems and the speakers' qualities and characterize the sentences' difficulty. Initially, a list of the 100 sentences and the 75 speakers were given as input to the synthesis module. It resulted in a set of 7500 synthesized audios, each one produced from a given sentence and speaker. Afterward, each synthesized speech was transcribed by each ASR system in the pool. Finally, a report with all transcribed sentences was generated. Given the transcribed sentences and the original ones, the $WAcc$ was calculated. Then, the computed $WAcc$ values were stored in the response matrix, which was the input to our IRT model. Finally, the IRT identifies the latent structure of the responses to estimate sentences' difficulties and discriminations, ASR systems' abilities and also speakers' qualities over the two levels of execution.

The results showed that the difficulty of the synthesized speeches and the transcription accuracy rates have a negative and very strong correlation. It indicates that the most difficult synthesized speeches are more likely to have a low transcription accuracy by the ASR systems. Moreover, we noticed that the highest is the difficulty of a sentence, the harder it will be to a speaker synthesize it. Furthermore, we observed that the higher the speaker's quality, it is supposed to synthesize sentences with less difficulty. So it seems this methodology could be helpful in a new ASR testing process. Suppose that we want to test an ASR system that is under development, which requires easy sentences at the beginning and more difficult sentences when the system is getting more robust. By adopting our proposal, we can rank the sentences according to the difficulty level and then different test data can be used at different stages of the testing process.

[1] Universidade Federal de Pernambuco, Recife (PE), Brasil, email: cso2@cin.ufpe.br

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 41–42, Santiago de Compostela, Spain, August 29-30, 2020.

The current case study conducted in our research provided interesting insights about how to evaluate ASR systems. The proposed methodology is promising, which motivate us to further investigate different IRT techniques and to perform new case studies. As the next step, we aim to perform a case study in a real scenario of ASR testing in partnership with Motorola Mobility.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Dr. Ricardo B. C. Prudêncio, for all support and guidance. We had the following paper accepted in the ECAI main conference: Chaina S. Oliveira, Caio C. A. Tenório, and Ricardo B. C. Prudêncio, 'Item response theory to estimate the latent ability of speech synthesizers', in 24th European Conference on Artificial Intelligence - ECAI (2020). This work was supported by CAPES, CNPq and FACEPE (Brazilian funding agencies) and Motorola Mobility.

REFERENCES

[1] Yu Chen, Telmo Silva Filho, Ricardo B. C. Prudêncio, Tom Diethe, and Peter Flach, 'β^3-irt: A new item response model and its applications', in *Proceedings of Machine Learning Research*, volume 89, pp. 1013–1021, (2019).

[2] Rafael Jaime De Ayala, *The theory and practice of item response theory*, Guilford Publications, 2013.

[3] Susan E. Embretson. and Steven P. Reise, *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Inc., 2000.

[4] Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo, 'Making sense of item response theory in machine learning', in *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pp. 1140–1148. IOS Press, (2016).

[5] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo, 'Item response theory in ai: Analysing machine learning classifiers at the instance level', *Artificial Intelligence*, **271**, 18–42, (2019).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 41–42, Santiago de Compostela, Spain, August 29-30, 2020.

Hate Speech Analysis via Argumentation Schemes Mining

Damian Furman [1]

Abstract. In this work we delineate Damian Furman's research objectives to be achieved in his doctoral studies in the next four years. The proposal is the formalization and development of a framework for characterizing hate speech in social media as particular instances of argumentative structures. This characterization would allow to automatically produce adequate and coherent counter-narratives.

1 RESEARCH PROBLEM BEING INVESTIGATED

Our work focuses on argument mining for social media for specific topics of discussion. In particular we are currently focusing on hate speech. Hate speech refers to statements that attack particular groups of people based on ethnicity, gender, religion, sexual orientation, or any other demographic category. It is very common to encounter such demonstrations in social media platforms, like Facebook and Twitter, as it can be a means for spreading violence and prejudice. Therefore it is crucial to being able to automatically identify such statements as they appear in order to moderate such forums. Though it is true that in many cases this expressions of hate resume basically to name calling or plain insults, we have find in several available data sets [1, 2] that also many of the statements classified as hate speech maintain certain argumentative form (mostly based on fallacies or wrong assumptions) that resemble closely the structure of Argumentation Schemes defines by Walton in [8, 7]. Therefore, our research tries to go a bit further in such direction, not only identifying potential posts carrying hate speech, but also being able to identify the argumentative structure behind them. Being able to identify the structure can tell us what the premises and assumptions behind the speech are, this is very useful as we can in turn analyze them to provide logically coherent responses and counter arguments and, ideally, being able to change social prejudices through the implementation of adequate social policies. This same idea can be applied to digest arguments for several other topics of interest (racism, climate change, legalization of abortion, etc.) as the fuel to produce automated production of counter-narratives. Studying this problem over social networks also provides hundred of thousands of unlabeled examples with extra information (like knowing that some example was posted in reply to some other example) that may also be used to improve the task of finding counter-narratives.

2 SPECIFIC CONTRIBUTIONS MADE

We have defined a set of four argumentation structures that are either adaptations or special cases of Walton Argumentation Schemes [8, 7] to account for common arguments in hate speech. Informally, the four schemes proposed so far are the following:

- Argument from generalization: generalizes particular cases of wrongdoing to a whole collective.
- Argument from fear and danger: appeals to the population fear and danger (and potential bad consequences) to argue against a person or a collective. We also identified related structures that involve the acts or doing of people of social influence or in a position of power or authority.
- Argument from division: argues that something that is true for the whole is also true for the parts of the whole
- Argument from statistical inference: uses statistical information to establish a position against a collective.

For each of these schemes we have identified the parts that bring about the argumentative statement, that is, premises, explicit and implicit assumptions, claims, etc. Consider the case of argument from fear and danger; the general structure of such schema would be the following:
Premise 1: Person or collective P produces A
Premise 2: If A occurs then B occurs
Implicit assumption: B is bad/dangerous
Claim: Degrade o demote P.

Now consider the following tweet extracted from [2]:

> Canada's @JustinTrudeau appointed to his cabinet immigrants from such cultures. One is in charge of immigration & refugee affairs - license to flood Canada with own kind. The other, Education Minister, tried to sneak into new Education code female genital mutilation as a right!

P: @JustinTradeau
A: appointed immigrants
B: flood canada with own kind - code female genital mutilation as a right.
Claim: demote P (the conclusion is implicit by the tone of the statement)

We are currently studying the prevalence of these schemes in hate speech corpora [2] and corpora for counter-narratives [3, 6].

3 DIRECTIONS FOR THE REMAINING WORK

Here is a high level description of the proposed plan of work:

- Manual annotation of the proposed schemes in the reference corpora [2, 3, 6]. The annotation will include a label that indicates to which of the schemes the statement belongs to and labels identifying the corresponding parts of the structure.

[1] Universidad de Buenos Aires, Argentina, email: damian.a.furman@gmail.com

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 43–44, Santiago de Compostela, Spain, August 29-30, 2020.

- These annotated data sets will serve as training and validating data for a classifier design to identify the targeted schemes and their components.
- Finally, we integrate the results of the classification in a Natural Language Generation (NLG) approach to produce counter-narratives.
- Empirical evaluation to demonstrate the value of the proposal will be carried out, which will include both the classification and counter-narrative generator tools.

4 STATE OF THE ART

Most of the work for social media (sentiment analysis, topic identification, etc.) and in particular hate speech analysis is usually addressed as a discrete or even binary classification problem, oversimplifying the target. We believe that a fine graded approach that involves semantics understanding of argumentative structures is needed in order to generate analysis and decision making support tools that can effectively help in mitigating undesired consequences of social media behavior. Furthermore, most of the proposals follow an approach of end-to-end neural architectures, with no insights in the inner workings of the phenomena (arguments, linguistic issues), which are also difficult to integrate with knowledge-rich approaches.

This sub-symbolic treatment of natural language processing has severe limits. When applied to argument mining, for instance, analyzes are shallow without discovery of argumentative sub-elements. Other crucial nuances of human expression are lost, even something as elemental as negations, while complex one such as the use irony remain a utopia.

The core of our proposal is to integrate finer-grained analysis of arguments to aid neural approaches to generate counter narratives for hate speech, at any of the stages of the solution: analysis of hate speech to identify arguments, identifying components of the arguments, finding a pre-generated counter-argument, generating the counter-argument via natural language generation techniques.

We believe it will be successful as this proposal rests on the bases of strong theoretical and practical evidence. On one hand, argument mining has been shown to be more successful when targeting a limited number of schemes [4]. Furthermore, it has been shown recently that it is feasible to successfully generate counter-narratives for hate speech [3, 6]. Finally, there exists a considerable amount of high quality data sets that can be used to prove our hypothesis; as they do not include argumentative annotation it is upon us to produce an adequate annotated corpora. The most significant difference with existing approaches to hate speech analysis is that our proposal aims for a symbolic representation and reasoning framework where specific pieces of knowledge that reveal the semantic of such statements can be identified automatically. The clear advantages of such frameworks are related to model interpretability and the possibility to being able to explain not just the computational task itself (for instance, why is this statement classified as hate speech) but also provide insights on the nature of the phenomenon (how and why people reach to such negative conclusions and actions).

We plan to evaluate our framework quantitatively, comparing the results to a human-generated gold standard [5] [9]. From a qualitative perspective we plan to compare our results to human performance both from crowds (for instance via a twitter bot) and from experts (for instance from NGOs [6]).

REFERENCES

[1] *https://github.com/marcoguerini/CONAN.*

[2] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti, 'SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter', in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 54–63, Minneapolis, Minnesota, USA, (June 2019). Association for Computational Linguistics.

[3] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini, 'CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2819–2829. Association for Computational Linguistics, (2019).

[4] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein, 'Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 386–396. Association for Computational Linguistics, (2018).

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 'Bleu: a method for automatic evaluation of machine translation', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, (July 2002). Association for Computational Linguistics.

[6] Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. Generating counter narratives against online hate speech: Data and strategies, 2020.

[7] Douglas Walton, 'Justification of argumentation schemes', *The Australasian Journal of Logic*, **3**, (2005).

[8] Douglas N. Walton, *Argumentation Schemes for Presumptive Reasoning*, L. Erlbaum Associates, 1996.

[9] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi, 'Bertscore: Evaluating text generation with BERT', *CoRR*, **abs/1904.09675**, (2019).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 43–44, Santiago de Compostela, Spain, August 29-30, 2020.

A framework for the automatic generation of natural language descriptions of processes

Yago Fontenla-Seco[1]

Abstract. Processes constitute a useful way of representing and structuring the activities that take place in organization information systems from almost any domain. However, understanding process models and their related statistics is not a trivial task, as complex representations of process structures and beahaviour appear when processes are implemented in real life. Often, only experts in the process field are capable of understanding them in detail, making experts on the domain the process takes place rely on the discourse or explanation this process analysts can provide them. This predoctoral project presents an approach for the automatic generation of natural language descriptions of processes and aims to create and validate a framework and a system that generates said descriptions.

1 INTRODUCTION

As every day more event data is being produced and stored by the execution of processes, it is necessary to provide organizations with tools capable of processing such vast amounts of data and extracting the valuable knowledge hidden in it. Process mining goal is to exploit that event data to extract valuable, process related information in a meaningful way. This information can be used to provide insights, determine performance, detect, and identify bottlenecks, anticipate problems, streamline and improve processes, etc. [9].

Process data is usually presented as process models (in a great variety of notations [9]) that represent in a graphical manner the activities that take place in a process as well as the dependencies between them. Other properties of the process can be included in the model such as temporal properties, the use and creation of data, decisions and resources involved in the execution of the process, etc.

The other most common way to communicate this information to users apart from process models are visual analytics (or data visualization), as they are commonly used when providing advanced analytics [9]. Visual analytics or data visualization focuses on the analysis of large amounts of data relying on human capabilities to understand and identify valuable information on data in form of graphs, dashboards, scorecards, etc. However, process models are often very complex, and this kind of visual analytics are usually quite difficult to be understood by a user on the organization or expert on the domain the process takes place, as a deep knowledge of processes modeling and analytics is required.

Parallel to data visualization, in the Natural Language Generation (NLG) and Fuzzy Logic fields, different methods for generating insights on data through natural language have been under development. Through different techniques, the aim is to provide users with natural language texts that capture or summarize the most characteristic aspects of the data that is being described. This information can be easily consumed by users, as natural language is the inherent way of communicating for humans, therefore it does not rely on their capabilities to identify or understand patterns, trends, etc. from visual representations.

In this context, natural language descriptions can be a good approach to enhance the understanding and description of processes and its analytics, as research shows that expertise is required to understand graphical information [6] and it has already been suggested that experts can take better decisions from textual summaries than from graphical displays [3].

This doctoral project aims on constructing a framework for the automatic generation of natural language descriptions of processes.

2 STATE OF THE ART

The generation of natural language texts from data is a task which originated within the NLG field. Particularly, the generation of natural language descriptions over data has been traditionally a task tackled by the Data-to-text (D2T) [7] research community. Parallel to NLG and D2T systems, other paradigms emerge from the fuzzy logic realm, for modelling and managing uncertainty in natural language. These paradigms use the concepts of linguistic summary and protoform [12, 13, 14, 15, 16] which aim on providing data summaries involving linguistic terms with some degree of uncertainty or ambiguity present on them.

In the process mining field, these two approaches can be each one paired with a research line. Following the first approach, focused on the control-flow perspective of a process [9], Henrik Leopold et al. in [5, 4] and with Han van der Aa in [8] try to provide a way to support process model validation and inconsistency detection using NLG and NLP techniques to generate natural language descriptions of process models. However, this approach only describes the control-flow perspective of a process, via its model. It does not use event logs neither process models extracted from event logs, so all the execution aspects of a process (activity frequencies, temporal distances, etc.) are missing i.e. the other three process perspectives described in [9].

The later approach, taken by Wilbik and Dijkman [10, 11, 1], applies the concepts of linguistic summaries and protoforms on event logs, focused on the case perspective of processes [9]. They propose a series of protoforms over sequences of activities and cases, able to describe different properties of cases e.g.: throuput time, frequency of a sequence... However, they only focus on cases, without taking in count the process structure (model) or the meaning of the activities of the process. To tackle the problem of the number of summaries generated, they cluster sequence descriptions using sequence-based

[1] University of Santiago de Compostela, Spain, email: yago.fontenla.seco@usc.es

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 45–46, Santiago de Compostela, Spain, August 29-30, 2020.

distances (like Levenshtein) that do not take in count the meaning of an activity on a process. This leads to grouping sequences with a high likelihood sequence-wise but that can be very different from the process domain viewpoint. This approach therefore lacks the genuine aspects of process mining and process analytics as well as the domain of the analyzed process.

Once analyzed the state of the art, we can see no approach has gone further than describing structural aspects of a process (process models) or using protoforms to describe sequences or cases from event logs, without the combination of structure and quantitive information and lacking genuine process elements such as temporal relations, process analytics and indicators based on activities, traces, etc.

3 CURRENT STATE AND PRELIMINARY RESULTS

At fist, only protoforms where used. Following a linguistic descriptions of data approach similar to the one proposed by Wilbik and Dijkman, however, using the process model and related statistics extracted from the event log rather than just the information present on the log. This made our approach more valid from the process mining and analysis viewpoint.

At this moment, a two-stage architecture (data interpretation and realization) based on the D2T architecture is implemented. This architecture is simpler than the full NLG architecture as some of the steps are manually defined parting from domain expert knowledge. It still uses fuzzy quantified protoforms in the data interpretation phase, as they are usefull for managing uncertainty and extracting relevant information from input data sets. Then, this protoforms are realized with a combination of template-based realization (the main structure of the description) and the SimpleNLG [2] realization engine.

4 FUTURE WORK

Future work includes the developement of a complete NLG system rather than a system that relies on template-based realization. This will give the framework more flexibility when generating descriptions as well as the capability to adapt to different processes and process domains without the need develop a new set of templates for each of them. A complete NLG system should also be capable of generating more syntactically complex descriptions and description aggregation, which give place to more human-like texts than those generated through templates, further helping on a better understanding of them. This system will be (the current implementation is being) validated over real life processes.

ACKNOWLEDGEMENTS

This doctoral project is supervised by Dr. Alberto Bugarn and Dr. Manuel Lama. Supported by grant PRE2018-083719 from the Spanish Ministry of Science, Innovation and Universities, aligned with the research project: Aportando inteligencia a lso procesos de negocio mediante soft computing en escenarios de datos masivos (TIN2017-84796-C2-1-R).

REFERENCES

[1] Remco M. Dijkman and Anna Wilbik, 'Linguistic summarization of event logs - A practical approach', *Inf. Syst.*, **67**, 114–125, (2017).

[2] Albert Gatt and Ehud Reiter, 'SimpleNLG: A realisation engine for practical applications', *Proceedings of the 12th European Workshop on Natural Language Generation, ENLG 2009*, (March), 90–93, (2009).

[3] Anna S. Law, Yvonne Freer, Jim Hunter, Robert H. Logie, Neil McIntosh, and John Quinn, 'A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit', *Journal of Clinical Monitoring and Computing*, **19**(3), 183–194, (2005).

[4] H. Leopold, J. Mendling, and A. Polyvyanyy, 'Supporting process model validation through natural language generation', *IEEE Transactions on Software Engineering*, **40**(8), 818–840, (Aug 2014).

[5] Henrik Leopold, Jan Mendling, and Artem Polyvyanyy, 'Generating natural language texts from business process models', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **7328 LNCS**, 64–79, (2012).

[6] Marian Petre, 'Why looking isn't always seeing: Readership skills and graphical programming.', *Commun. ACM*, **38**, 33–44, (06 1995).

[7] Ehud Reiter, 'An Architecture for Data-to-Text Systems', in *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pp. 97–104, USA, (2007). Association for Computational Linguistics.

[8] Han van der Aa, Henrik Leopold, and Hajo A. Reijers, 'Detecting inconsistencies between process models and textual descriptions', in *Business Process Management*, eds., Hamid Reza Motahari-Nezhad, Jan Recker, and Matthias Weidlich, pp. 90–105, Cham, (2015). Springer International Publishing.

[9] W.M.P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, volume 136, 2011.

[10] Anna Wilbik and Remco M. Dijkman, 'Linguistic summaries of process data', *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–7, (2015).

[11] Anna Wilbik and Remco M. Dijkman, 'On the generation of useful linguistic summaries of sequences', in *2016 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2016, Vancouver, BC, Canada, July 24-29, 2016*, pp. 555–562, (2016).

[12] Ronald R. Yager, 'A new approach to the summarization of data', *Information Sciences*, **28**(1), 69–86, (1982).

[13] L A Zadeh, 'Fuzzy logic = computing with words', *IEEE Transactions on Fuzzy Systems*, **4**(2), 103–111, (may 1996).

[14] L. A. Zadeh, 'From computing with numbers to computing with words. from manipulation of measurements to manipulation of perceptions', *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, **46**(1), 105–119, (Jan 1999).

[15] L A Zadeh, 'A prototype-centered approach to adding deduction capability to search engines-the concept of protoform', in *2002 Annual Meeting of the North American Fuzzy Information Processing Society Proceedings. NAFIPS-FLINT 2002 (Cat. No. 02TH8622)*, pp. 523–525, (2002).

[16] Lotfi A. Zadeh, 'Toward a generalized theory of uncertainty (GTU)- An outline', *Information Sciences*, **172**(1-2), 1–40, (2005).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 45–46, Santiago de Compostela, Spain, August 29-30, 2020.

A Model for adopting the omnichannel strategy from a Context-aware computing and Natural Language Processing approach

Carlos López[1]

Abstract. The omnichannel approach allows the organizations for offering to customers a more seamless, consistent and integrated cross-channel experience. We can find in the state of the art some models for supporting the management task of multichannel contact centers, but such models lack a mechanism for allowing the integration among channels that enable the companies to deliver the omnichannel experience. In this research, the omnichannel problem is addressed from the design of a system based on the context-aware computing approach with the Natural Language processing perspective.

1 PROBLEM STATEMENT

Modeling contact center has constituted a research field with some proposals over time, whose goals has been centered on providing tools in order to facilitate its administration. The emergence of service management strategies like the omnichannel approach, make necessary the design of such models by considering both the new communication channels and the omnichannel approach goals itself. In the contact center domain, the omnichannel approach is related to offer "a seamless and integrated environment for modern customer experience (CX) through integrated channels, which allows agents to work on a better interface and to use a richer set of customer and service data" [1].

We can find some models in the state of the art for measuring and predicting the contact center performance as a way to determine the necessary resources to satisfy certain demand and reach a certain performance level. The physical infrastructure, the staff, and the budget are some of the resources considering in these models. Some of these models are based in statistical analysis techniques [2][3], in the predictive analytics [4][5], the optimization [6], and the simulation [7]. These models make important contributions to the contact center management, buy they are focused on the contact center operative challenges, like staff, scheduling, physical resource assignment, and the forecasting of the interactions volume. However, such models are not supposed to address the omnichannel approach in the contact center domain and their respective performance measure.

In addition, even though we have models for contact centers that recognize the need for adopting an omnichannel strategy [7], this model faces some challenges facing in this domain. Some of such challenges are the data acquisition from the channels, the data privacy, the customer identification and authentication trough the channels, the collected data synchronization and the channel integration. Such challenges add limitations to the previously described models related the adoption and management of the omnichannel approach in the contact center domain. They also reveal the lack of a method/model for allowing the adoption of the omnichannel approach in this domain and consider the necessary tools and techniques for the processing of relevant information to the omnichannel approach implementation. Such information can be extracted and reasoned from the linguistic content of the interaction established between the users and the contact center agents.

[1] Computer and Dicision Science department, Universidad Nacional de Colombia, Colombia, email: calopezj@unal.edu.co

2 THEORETICAL FRAMEWORK

2.1 Context-aware computing

"A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task" [8]. Some of the tasks included in the research field correspondent to the context-aware computing scope are what to sense, how to acquire the information and reasoning to that information to infer the context of a user [8]. "Developing such context-aware applications is inherently complex and hence should be supported by adequate context information modeling and reasoning techniques" [8]. The author define context as "any information that can be used to characterize the situation of an entity," and they define an entity as "a person, place, or object can become entity which is considered applicable to the interaction between a user and an application, including the user and applications themselves." On this meaning, designing a system that facilities an omnichannel strategy delivery is tied to the need of know and recognize the user context according with the definition of the omnichannel concept. The following summarized steps to convert contextual information into smart actions are presented in [8]: create computational models for context acquisition with respect to tasks; generalization of contextual characteristics based on the tasks; aggregation of gathered information to generate an abstraction of context; selection of the algorithm or a set of algorithms to use; context recognition and inference derivation; and conclusion derivation and user assistance with most suitable decisions.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 47–48, Santiago de Compostela, Spain, August 29-30, 2020.

2.2 Natural language processing

"Natural language processing (NLP) is a theory-motivated range of computational techniques for the automatic analysis and representation of human language" [9]. We are particularly interested in the summarization in this research, witch is a technique to "reduce the size of the document while preserving the meaning" [10]. The main summarization techniques are the extractive and the abstractive techniques. "Extractive summarization is to find out the most salient sentences from the text by considering the statistical features and then arranging the extracted sentences to create the summary" [10], whereas abstractive summarization is "a technique in which the summary is generated by generating novel sentences by either rephrasing or using the new words, instead of simply extracting the important sentences" [10]. The internal representation of the text is created by analyzing the semantic content of the text, and new sentences are generated from the original text by using deep learning and reasoning. The main approach to abstraction techniques is based on: Structure-Based (tree based, template based, ontology based, lead and body phrase, graph based and rule based), Semantic-Based (predicate argument based, semantic graph based, information item based, multimodal), Deep Learning with Neural Networks and Discourse and Rhetoric Structure Based [10].

3 PROPOSED SOLUTION

The general objective of this research is building a model for allowing the adoption of the omnichannel strategy in the contact center domain, based on the context-aware computing and the natural language processing approach. We hope to achieve this goal through the following specific goals:

1. Characterizing the user context from a knowledge representation of the contact center domain.
2. Performing the recognition of the user context from the extraction and processing of the conversational content of the interactions the user engages with the contact center.
3. Developing a channel integration model based on the definition of a common language for representing the user context and shared across different channels; in this way, each channel can have the necessary information to deliver the omnichannel experience.
4. Validating the model by using a case study for implementing the model in a real contact center.

The methodology used in this research is based in the paradigm Design Science In information Research [11], based on a conceptual framework and a set of guidelines which are the basis for the next stages defined for this work: Exploration (literature review, problem identification, statement of research objectives, determination of research contributions), Analysis (definition of the design evaluation criteria and key performance indicators of the model, compilation of the linguistic corpus, characterization and modeling of the user context in the contact center domain), Development (extraction and representation of the knowledge of the contact center domain from the compiled linguistic corpus, design of the method for extracting the user context from the interactions, modeling the channel integration, integration of the models built for consolidating the omnichannel approach model), Validation (case study) and Communication (scientific writing and result divulgation).

4 EXPECTED RESULTS

This research will provide a model for allowing the adoption of the omnichannel strategy, benefiting the organization who desire deliver a more seamless experience to their customers, reaching their loyalty, confidence, and engagement. Compiling a linguistic corpus, we will extract the contact center ontology, based on the conversational content of the interactions established in the different channels, such as social media, email, voice calls, chats, etc. We will represent the user context with this ontology and then we will define the service provided by the channel like a context-aware system. This system will have the ability of extract and reason the contextual information from the conversational content of each interaction, where each channel will have the awareness of the user context, and the capability of share this user context with others channels, delivering the omnichannel experience.

ACKNOWLEDGMENT

I would like to thank to my thesis directors Carlos Zapata and Bell Manrique for supporting this initial phase of my PhD degree.

REFERENCES

[1] R. Picek, D. Peras, and R. Mekovec, 'Opportunities and challenges of applying omnichannel approach to contact center', *2018 4th International Conference on Information Management*, 231–235, (2018).

[2] J. Huerta, 'A model for contact center analysis and simulation', *Winter Simulation Conference, Simulation Conference, 2007 Winter*. 2259–2265, (2007).

[3] A. Bastianin, M. Galeotti and M. Manera, 'Statistical and economic evaluation of time series models for forecasting arrivals at call centers', *Empirical Economics*, 1–33, (2018).

[4] S. Moazeni and R. Andrade, 'A Data-Driven Approach to Predict an Individual Customer's Call Arrival in Multichannel Customer Support Centers', *IEEE International Congress on Big Data, BigData Congress - Part of the IEEE World Congress on Services*, 66–73, (2018).

[5] L. Xu, X. Hu, X. L. Wang, and G. Huang, 'Forecasting of Intraday Interval Arrivals for Small and Medium Sized Call Centers with Emergencies', *Procedia CIRP*, 56(71571101), 456–460, (2016).

[6] N. Ilk, M. Brusco and P. Goes, 'Workforce management in omnichannel service centers with heterogeneous channel response urgencies', *Decision Support Systems*, 105, 13–23, (2018).

[7] P. Liston, J. Byrne, O. Keogh and P.J. Byrney, 'Beyond calls: Modeling the connection center', *2017 Winter Simulation Conference*, 4220–4227, (2017).

[8] P. S. Gandodhar and S. M. Chaware. 'Context aware computing systems: A survey'. *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, 605–608, (2019).

[9] T. Young and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing", *IEEE Computational Intelligence Magazine*, 13, 55–75, (2018).

[10] S. Gupta and S. K. Gupta. 'Abstractive summarization: An overview of the state of the art', *Expert Systems with Applications*, 121, 49–65, (2019).

[11] A. Hevner, S. March and S. Ram, 'Design Science in Information Systems Research', *MIS Quarterly*, 28(1), 75–105, (2004).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 47–48, Santiago de Compostela, Spain, August 29-30, 2020.

Structures and Infrastructures around AI: Unbiasedness, Politics and Metrics in Data-driven Socio-technical Systems

Gunay Kazimzade[1]

Abstract. The purposed research project aims for interdisciplinary analysis of structures and infrastructures around Artificial Intelligence-based systems and discussing the conceptional and technical concepts around the problem of bias and discrimination in those AI-based systems.

1 EXTENDED ABSTRACT

Generally, biases occur in all forms of discrimination in our society, for example, on a political, cultural, financial, or sexual level. These are again manifested in the data sets collected as well as the structures and infrastructures around the data, technology, and society, and thus represent social standards and decision-making behavior in particular data points.

In the past decade, Artificial Intelligence (AI)-based systems are increasingly regulating major critical areas of human life. Predictive technologies, scoring systems, social ranking algorithms are being implemented in socio-technical systems at a state level [12]. Nevertheless, there is a significant number of cases where data-driven systems are producing discriminatory decisions for characteristics based on gender, age group, dis/ability status, and ethnic descent groups [1,2,12].

These issues bring a new set of questions involving a variety of entities, stakeholders, legal, and economic structures concerned and not-ready for the discriminatory challenges of data-driven decision-making [8].

The traditional approach towards the Ethics of AI and "unbiasedness" is generally concentrated on the so-called "AI pipeline", reflecting on detecting, measuring, and mitigating biases in concrete "layers" of AI system development. Primary research contributes to the problem of biased datasets, inclusion in design, biased machine learning models with the layered focus on data cleaning, data manipulation, design, modeling, and implementation.

However, recent research findings around the topic of bias and discrimination in AI highlight that, even if we can make datasets and algorithms measurably "unbiased", all other issues of the socio-technical systems into which AI algorithms are embedded need looking into, might be severely outdated in terms of monolithic, controllable systems [3,4,5,6,7,13]. Some of the much-investigated questions around "unfairness" are just the tip of the iceberg, and by solving one problem, we do forget to solve the more significant other problems [8,10,15].

[1] Faculty IV, Technical University of Berlin, Germany, email: gunay.kazimzade@tu-berlin.de

The creation of data-driven systems does not happen in a vacuum but profoundly impacted by human decisions, values, priorities, culture, and prejudices [8,14,15]. Discriminatory practices experienced by AI are the reflection of power asymmetries, workplace structures, and overall fundamental politics and power distribution in society [16].

With these concerns in our mind, we found value on concentrating our focus on politics and metrics around AI-based systems and investigating the interplay between society and technology (AI) considering biased values, motivation, and standards in the creation of those systems. In our previous study, we concentrated on data creation and transformation structure, and the paper "Biased Priorities, Biased Outcomes: Three Recommendations for Ethics-oriented Data Annotation Practices" was published in AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. The study aimed to analyze the relationship between biased data-driven outcomes and practices of data annotation for vision models, by placing them in the context of market economy and argued about the prevalence of market-oriented values over socially responsible approaches. We have investigated the goals prioritized by decision-makers at the data creation structure and the way these priorities correlate with data-related bias issues and come up with three recommendations for ethics-oriented data-annotation practices [9].

The outcomes of this study merged the most common prioritized patterns as *profit, standardization, and opacity* around the creation and annotation of image data sets [9]. Moreover, these results contributed to extracting three further aspects, namely *transparency, education, and regulations* that should be prioritized by companies and policymakers, if fighting biased AI is to be taken seriously [9].

Being extremely concerned about the implementation and real-world effect of the recommendations made based on our investigations, as the next step, we are aiming to work closely with the impact-sourcing companies, as well as other stakeholders in the data-creation structure for practical approaches towards ethics-oriented data creation practices. Nevertheless, there are several further questions we are aiming to address in the next stages of the current research project.

- How can a social consensus be reached on new standards for intelligent systems to place their development as a democracy-friendly, ethical, and unbiased process under the responsibility of all individuals?
- How can the hierarchy between the system decision maker and system developer be deconstructed by

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 49–50, Santiago de Compostela, Spain, August 29-30, 2020.

unequally distributed knowledge about digital technologies?

- How standards and ethical guidelines can promote non-discriminating decision-making in AI systems?

With this questions in our mind, we are planning to concentrate on a specific pilot data-driven ranking, decision and recommendation systems implemented in various states and authorities (e.g., credit scoring, unemployment support, education) to be able to analyze the given questions with respect to technical and social implications as well as the conceptual principles around those systems. In the next stage of our analysis, we are willing to dive into the area of data representation in socio-technical systems and visioning for conceptional and technical analysis around the problem of bias in those AI-based systems.

We are aiming to apply further mixed methods, including metrics of quality engineering, literature research, discourse analysis, to address the introduced research directions. Insights from critical data studies [1,2] and research on fairness, accountability, and transparency in algorithmic systems [10, 11] are the main inspirations for the current research project.

ACKNOWLEDGEMENTS

Funded by the German Federal Ministry of Education and Research (BMBF) -NR 16DII113.
We thank our advisors at Weizenbaum Institute, especially Prof. Dr. Bettina Berendt, who provided helpful comments and valuable suggestions to mature the ideas for this research project, Milagros Miceli who contributed to the previous studies and co-authored the papers, Martin Schüßler and Tianling Yang who contributed with the data collection and processing to the previous studies. Special thanks to our research group leader, Dr. Diana Serbanescu, for her continuous support of this project.

REFERENCES

[1] Cathy O'Neil. 2017. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (01 edition ed.). Penguin, London.

[2] danah boyd and Kate Crawford. 2012. Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15, 5 (June 2012), 662–679. DOI:https://doi.org/10.1080/1369118X.2012.678878

[3] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18,* 67–73. DOI:https://doi.org/10.1145/3278721.3278729

[4] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumee III, and Kate Crawford. 2018. Datasheets for Datasets. arXiv:1803.09010 [cs] (March 2018). *Retrieved September 12, 2019, from http://arxiv.org/abs/1803.09010*

[5] Ansgar Koene, Liz Dowthwaite, and Suchana Seth. 2018. IEEE P7003TM standard for algorithmic bias considerations: work in progress paper. *In Proceedings of the International Workshop on Software Fairness - FairWare* '18, 38–41. DOI:https://doi.org/10.1145/3194770.3194773

[6] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. SSRN *Electronic Journal (2016).* DOI:https://doi.org/10.2139/ssrn.2886526

[7] Kathleen H. Pine and Max Liboiron. 2015. The Politics of Measurement and Action. *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15),* 3147–3156. DOI:https://doi.org/10.1145/2702123.2702298

[8] West, Whittaker, and Crawford, "Discriminating Systems. Gender, Race, and Power in AI."

[9] Gunay Kazimzade and Milagros Miceli. 2020. Biased Priorities, Biased Outcomes: Three Recommendations for Ethics-oriented Data Annotation Practices. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 71. DOI:https://doi.org/10.1145/3375627.3375809*

[10] Crawford, Kate, and Jason Schultz. "AI SYSTEMS AS STATE ACTORS." *Columbia Law Review, vol. 119, no. 7, 2019, pp. 1941–1972.* JSTOR, www.jstor.org/stable/26810855. Accessed 10 Mar. 2020.

[11] Sandvig, Christian et al. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." (2014). G.H. Golub and C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 2nd edn., 1989.

[12] Niklas, Jędrzej. "PROFILING THE UNEMPLOYED IN POLAND : SOCIAL AND POLITICAL IMPLICATIONS OF ALGORITHMIC DECISION MAKING." (2015).

[13] Calderon et al. (2019): Calderon, Ania; Taber, Dan; Qu, Hong; Wen, Jeff: AI Blindspot: A Discovery Process for preventing, detecting, and mitigating bias in AI systems. 2019, URL: http://aiblindspot.media.mit.edu/ [Zugriff: 12.9.2019]

[14] Milagros Miceli. 2019. Data Labeling Work between Subjectivity and Standardization. A Grounded Theory Study. Unpublished master thesis, Humboldt-Universität zu Berlin.

[15] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for Datasets. arXiv:1803.09010 [cs] (March 2018). *Retrieved September 12,* 2019 from http://arxiv.org/abs/1803.09010

[16] Rob Kitchin. 2014. The data revolution: big data, open data, data infrastructures & their consequences. *SAGE Publications, Los Angeles, California.*

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 49–50, Santiago de Compostela, Spain, August 29-30, 2020.

Accountability and Control over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight

Ilse Verdiesen [1]

Abstract. Accountability and responsibility are key concepts in the academic and societal debate on Autonomous Weapon Systems, but not yet operationalized. Accountability is a notion of *backward-looking* responsibility and control can be viewed from different perspectives. I propose a Framework for Comprehensive Human Oversight that connects the engineering, socio-technical and governance perspective of control. By this I aim to broaden the view on the control over AWS which may ensure solid controllability and accountability for the behaviour of AWS. Comparing the CHO Framework of the deployment of current weapon systems to that of AWS reveals two gaps in the control mechanisms. I have identified three first options for the development of such a mechanism. In future work, I aim to design a mechanism based on the Glass-box approach to fill these gaps.

1 INTRODUCTION

Accountability and responsibility are key concepts in the academic and societal debate on Autonomous Weapon Systems (AWS). However, these notions are mentioned as high-level principles [1] but not yet operationalized. The concept of Meaningful Human Control (MHC) is often seen as requirement for the deployment of AWS [2-4], but this term is not-well defined and quantifying the level of control needed is hard [5]. Several scholars are working on defining the concept of MHC in Autonomous (Weapon) Systems [6-8]. Ekelhof [9] states that the relationship between the human operator and AWS is often used as reference to define MHC, but this is only one aspect of MHC. Mecacci and Santoni De Sio [7] take a wider conception of MHC and they show that the narrow focus of engineering control needs to be broadened to allow humans to directly influence the behaviour of autonomous systems in more than a narrow engineering notion of control.

However, this wider conception of the control loop does not incorporate the social institutional and design dimension at a governance level. The governance level is the most important level for oversight and needs to be added to the control loop, because accountability requires strong mechanisms in order to oversee, discuss and verify the behaviour of the system to check if its behaviour is aligned with human values and norms. Therefore, I am working on a Framework for Comprehensive Human Oversight (CHO) (figure 1) based on an engineering, socio-technical and governance perspective on control which may ensure solid controllability and accountability for the behaviour of AWS. By this I aim to broaden the view on the control over AWS and take a comprehensive approach. I am applying the Value-Sensitive Design (VSD) method [10] as research approach. The VSD is a three-partite approach for considering human values throughout the design process of technology.

The scientific contribution of my research is twofold in that it describes 1) the link between accountability and control and identifies the gap in control in the deployment of AWS and 2) a mechanism for control over AWS might also be applied to other AI fields to enhance transparency of decision-making by algorithms for Autonomous Systems, such as those for Autonomous Vehicles or in the medical domain. The societal contribution of my research is the application of a mechanism for control that allows a supervisor of an Autonomous (Weapon) System to monitor the actions of the system without having extensive knowledge of the internal workings of Autonomous (Weapon) Systems.

2 ACCOUNTABILITY AND CONTROL

Accountability is a notion of *backward-looking* responsibility [14] and is described by Bovens [15] as a mechanism of giving account of your actions by informing and explaining your decisions to others. According to Bovens [15] there is a fine line between accountability and control and he eloquently states: '*Accountability is a form of control, but not all forms of control are accountability mechanisms*.'

Control can be viewed from an *engineering*, a *socio-technical* and *governance* perspective which are based on the layers described by Van den Berg [16]. Meaningful Human Control does not always require more traditional forms of technical control. But accountability always requires strong mechanisms in order to oversee, discuss and verify the behaviour of the system to check if its behaviour is aligned with human values and norms. Therefore, I propose a Framework for CHO that connects the *engineering*, *socio-technical* and *governance* perspective of control.

2.1 CHO Framework

The CHO Framework (figure 1) consists of three horizontal layers that are based on the three-layered model of Van den Berg [16]. These layers can be linked to control perspectives described above. On the x-axis time is plotted which can be divided into three phases: 1) *before* deployment of a weapon, 2) *during* deployment of a weapon and 3) *after* deployment of a weapon. These phases are depicted by the vertical columns of the framework. The y-axis

[1] Delft University of Technology, The Netherlands, email: e.p.verdiesen@tudelft.nl

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 51–52, Santiago de Compostela, Spain, August 29-30, 2020.

describes the environment of the system which can range from internal to external. The *technical layer* describes the technical conditions required for the system to remain under control. The *socio-technical layer* describes the operators' psychological and motivational conditions required for the system to remain under control. The *governance layer* describes the political and institutional conditions and the oversight mechanisms required for the system to remain under control. The combination of layers and columns result in nine blocks that each contain a component of control in each phase and layer. The components of control are described in governance literature.

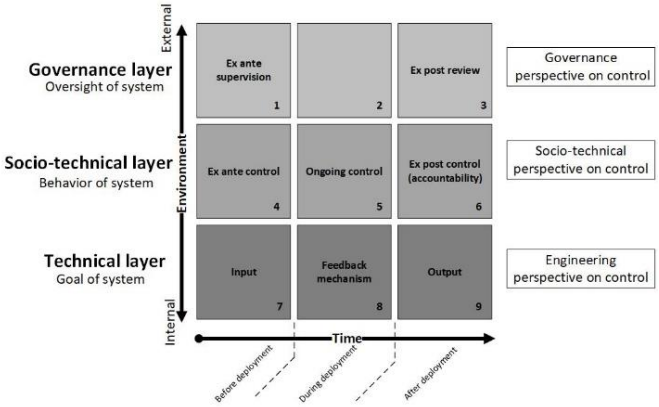

Figure 1 Comprehensive Human Oversight Framework

2.2 Gaps in CHO Framework

Comparing the CHO Framework of the deployment of current weapon systems to that of AWS reveals two gaps in the control mechanisms: 1) both frameworks lack a mechanism in block 2 that ensures oversight of a weapon during deployment, and 2) in the CHO Framework for AWS there is no ongoing control mechanism in block 5 to control the specific actions that the AWS takes to achieve its goal, because executive autonomy [17] inherently implies that the AWS is autonomous in setting its means to achieve its goal independently from the human operator.

3 FUTURE WORK

The CHO Framework for AWS shows two gaps and this raises the issue if this control framework is sufficient for control to be meaningful for the deployment of AWS. It seems that this is not the case and this deficiency indicates a need for an additional mechanism for the deployment of AWS. I have identified three first options for the development of such a mechanism: 1) a monitoring process in block 2 to ensure oversight of weapon system, 2) a mechanism in block 5 of the socio-technical layer, or 3) a mechanism in block 8 of technical layer to control the goal of the system. In future work, I aim to design a mechanism based on the Glass-box approach [11, 12]. The Glass-box approach has an interpretation stage, that derives requirements based on values and norms, and an observation stage that qualifies the behaviour of the system based on the rules as input and verifies the output. I will use the Colored Petri Nets (CPN) method as a computational mechanism to verify if the output of a system complies with the rules that were given as input [13].

3.1 Evaluation & Limitations

In the validation phase of my research I intend to evaluate the mechanism for CHO by running 2 or 3 scenarios to validate the architecture. The type of scenarios could entail both traditional physical AWS and cyber AWS.

The main challenge of my research approach lies in formalizing values, norms and requirements written in natural language and translating them in rules that can be verified by a logical program. The limitation of my work might be that I am conducting my research in the military domain and my findings might not be generalizable or applicable to other domains as I am studying a very specific field. Also, the formalization of values, norms and requirements into rules means that I will lose a lot of context that cannot be captured in logical formulization.

REFERENCES

[1] UN GGE LAWS, "Emerging Commonalities, Conclusions and Recommendations," 2018.

[2] T. K. Adams, "Future warfare and the decline of human decisionmaking," Parameters, vol. 31, no. 4, pp. 57-71, 2001.

[3] H. M. Roff, and R. Moyes, "Meaningful human control, artificial intelligence and autonomous weapons."

[4] K. Vignard, "The weaponization of increasingly autonomous technologies: considering how meaningful human control might move discussion forward," UNIDIR Resources, vol. 2, 2014.

[5] ICRC, Ethics and autonomous weapon systems: An ethical basis for human control?, International Committee of the Red Cross (ICRC), Geneva, 2018.

[6] M. Horowitz, and P. Scharre, Meaningful human control in weapon systems: a primer: Center for a New American Security, 2015.

[7] G. Mecacci, and F. Santoni De Sio, "Meaningful human control as reason-responsiveness: the case of dual-mode vehicles," Ethics and Information Technology, 2019.

[8] F. Santoni de Sio, and J. Van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," Frontiers in Robotics and AI, vol. 5, pp. 15, 2018.

[9] M. Ekelhof, "Moving beyond semantics on autonomous weapons: Meaningful human control in operation," Global Policy, vol. 10, no. 3, pp. 343-348, 2019.

[10] B. Friedman, P. Kahn, and A. Borning, "Value sensitive design: Theory and methods," University of Washington technical report, pp. 02-12, 2002.

[11] A. Aler Tubella, and V. Dignum, "The Glass Box Approach: Verifying Contextual Adherence to Values."

[12] A. A. Tubella, A. Theodorou, V. Dignum, and F. Dignum, "Governance by glass-box: implementing transparent moral bounds for AI behaviour," arXiv preprint arXiv:1905.04994, 2019.

[13] J. Jiang, H. Aldewereld, V. Dignum, and Y.-H. Tan, "Compliance checking of organizational interactions," ACM Transactions on Management Information Systems (TMIS), vol. 5, no. 4, pp. 1-24, 2014.

[14] I. Van de Poel, "The relation between forward-looking and backward-looking responsibility," Moral responsibility, pp. 37-52: Springer, 2011.

[15] M. Bovens, "Analysing and assessing accountability: A conceptual framework 1," European law journal, vol. 13, no. 4, pp. 447-468, 2007.

[16] J. Van den Berg, "Wat maakt cyber security anders dan informatiebeveiliging?," Magazine Nationale Veiligheid en Crisisbeheersing,(2) 2015, 2015.

[17] C. Castelfranchi, and R. Falcone, "From automaticity to autonomy: the frontier of artificial agents," Agent Autonomy, pp. 103-136: Springer, 2003.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 51–52, Santiago de Compostela, Spain, August 29-30, 2020.

Explainable AI for Intelligent Decision Support in Operations & Maintenance of Wind Turbines

Joyjit Chatterjee[1]

Abstract. As global efforts in transitioning to sustainable energy sources rise, wind energy has become a leading renewable energy resource. However, turbines are complex engineering systems and rely on effective operations & maintenance (O&M) to prevent catastrophic failures in sub-components (gearbox, generator, etc.). Wind turbines have multiple sensors embedded within their sub-components which regularly measure key internal and external parameters (generator bearing temperature, rotor speed, wind speed etc.) in the form of Supervisory Control & Data Acquisition (SCADA) data. While existing studies have focused on applying ML techniques towards anomaly prediction in turbines based on SCADA data, they have not been supported with transparent decisions, owing to the inherent black box nature of ML models. In this project, we aim to explore transparent and intelligent decision support in O&M of turbines, by predicting faults and providing human-intelligible maintenance strategies to avert and fix the underlying causes. We envisage that in contributing to explainable AI for the wind industry, our method would help make turbines more reliable, encouraging more organisations to switch to renewable energy sources for combating climate change.

1 INTRODUCTION

Condition based monitoring (CBM) has been of active interest to the wind industry, with the most popular approaches applying signal processing and numerical physics-based models [10]. Data-driven approaches have been also explored, with traditional ML algorithms including support vector machines, decision trees trees and probabilistic models being used for anomaly prediction [12, 1]. Some studies have utilised turbine power curves for identifying abnormalities in operation [5], but lack ability to provide component-level fault prediction (e.g. in gearbox). Deep learning algorithms have recently outperformed traditional ML methods in anomaly prediction in turbine operation, with the state-of-art studies utilising multi-layer perceptron and artificial neural networks [7]. The utilisation of more sophisticated architectures like recurrent neural nets has sadly been limited to a few studies applying long short-term memory models for power forecasting and fault diagnosis [8]. Some studies have proven effective in using computer vision for visual inspection via drones of external turbine sub-components (e.g. blades) applying convolutional neural networks [13], but are difficult to apply internally.

[1] University of Hull, United Kingdom, email: j.chatterjee-2018@hull.ac.uk

2 PROBLEM STATEMENT AND PROPOSED APPROACH

While existing studies have demonstrated significant advances in making more accurate power forecasts and anomaly prediction, they lack transparency to provide human-intelligible causes and maintenance actions, which is essential for engineers & technicians to consider for averting (or fixing) faults. This makes the wind turbine operators reluctant to adapt data-driven approaches widely, which we aim to address in this project. The more interesting information which has mostly been neglected includes unstructured data on historical alarms which have occurred in the turbine. These alarm records are stored as event descriptions for the faults, which are basically natural language phrases providing detailed information about the faults. In this project, given a sequence of continuous numeric SCADA input features, we aim to **(1)** Predict a fault type and generate the alarm event description, with its possible causes. **(2)** Generate a maintenance action message to avert/fix the fault.[2] The maintenance actions to fix failures can either be authored by a human-domain expert, or learnt from documents such as maintenance manuals and work orders. This is a data-to-text generation problem, wherein, deep learning and natural language generation (NLG) models have shown success in domains such as weather forecast generation [11], spatial navigation [9] etc. Owing to the sequential nature of inputs (continuous SCADA time-series) and sequential nature of outputs (predicted alarm messages and maintenance actions), we believe that models such as Seq2Seq [6] and Transformers[2] can provide transparent decisions beyond accurate predictions of faults. The components of the proposed approach are briefly outlined below (refer Figure 1):-

- **Stage (a): Alarm message generation module:** The first component of the system utilises a NLG model, which takes in the sequence of continuous SCADA features, and outputs the internal status of the turbine in the form of predicted alarm messages. The NLG model, such as Transformer provides prediction of likely occurrence of faults in advance [4] (to assist in preventive maintenance), while also estimating the potential causes of the fault (through the attention weights). [3]
- **Stage (b): Maintenance action generation module:** Considering a fault predicted in any turbine component, we propose utilisation of a second NLG model such as Transformer, which takes as input a sequence of likely causes of the fault in Stage (a) along

[2] We aim to develop a scalable system through transfer learning techniques, wherein, faults can be predicted in new domains (e.g. new wind farms which have not been in operation for long) without additional labelled training data.

[3] This component is trained using historical SCADA data labelled with corresponding alarm messages through supervised learning.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 53–54, Santiago de Compostela, Spain, August 29-30, 2020.

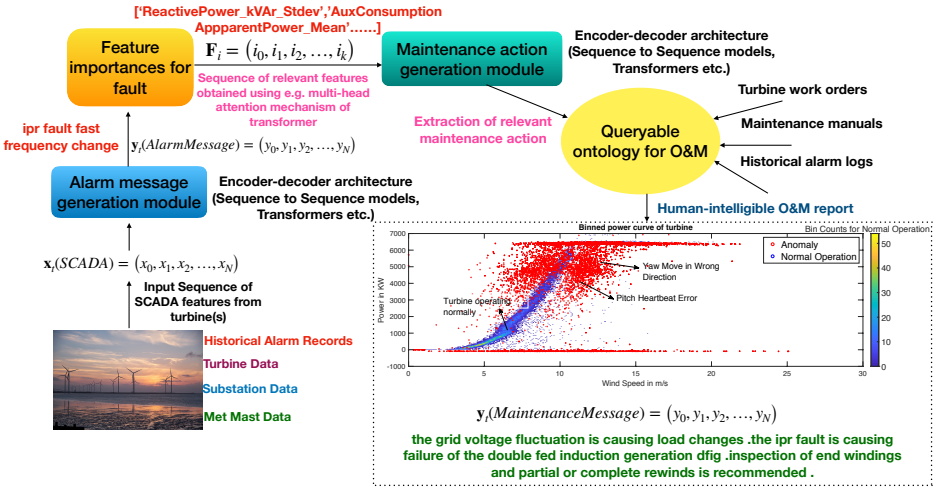

Figure 1. Our proposed intelligent decision support system for O&M of turbines.

with the identified alarm type, and generates the corresponding maintenance actions suitable for averting/fixing the fault. This is a content-selection problem, wherein, the most appropriate actions need to be selected from the available corpus [4].

3 PROGRESS AND RESEARCH PLAN

In the initial phase, we focused on obtaining SCADA data from an operational turbine [5], and its pre-processing. We established benchmarks on our dataset, and utilised various ML algorithms and deep learning techniques for comparison with existing work. At this stage, we developed a novel model utilising combination of a Long short-term memory recurrent neural network architecture for component-level predictions of faults and an XGBoost decision tree classifier to provide transparency to the black-box neural net. We also implemented transfer learning to port our model to an onshore wind farm, to predict faults without additional training data in a new domain [3]. Next, we extended our technique towards generating alarm messages and maintenance actions, utilising a dual-transformer NLG model for alarm type prediction and content selection [4]. At this point of submission, we are exploring development of a queryable ontology for O&M of turbines, by utilising maintenance manuals and other unstructured data. We are also considering the causal relationships in SCADA data to identify hidden relationships between features during various types of faults, through temporal causal graphs. Finally, we envisage that our approach will help facilitate intelligent decision support for the wind industry, by generating human-intelligible O&M reports in an accurate, scalable and transparent manner.

ACKNOWLEDGEMENTS

I am grateful to my PhD. Supervisor, Dr. Nina Dethlefs for the valuable guidance and support. Also, I would acknowledge ORE Catapult for providing turbine data through Platform for Operational Data.

[4] Content selection can be performed either through a collection of NLG templates authored by a domain-expert, or maintenance manuals etc.

[5] Special Acknowledgment: Platform for Operational Data (POD) Disseminated by ORE Catapult: https://pod.ore.catapult.org.uk

REFERENCES

[1] I. Abdallah, V. Dertimanis, H. Mylonas, K. Tatsis, E. Chatzi, N. Dervilis, K. Worden, and E. Maguire, 'Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data', in *Proceedings of the European Safety and Reliability Conference*, pp. 3053–3061, Trondheim, Norway, (June 2018).

[2] Ashish Vaswani et al., 'Attention is all you need', in *Advances in Neural Information Processing Systems 30*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998–6008, Curran Associates, Inc., (2017).

[3] Joyjit Chatterjee and Nina Dethlefs, 'Deep learning with knowledge transfer for explainable anomaly prediction in wind turbines', *Wind Energy*, **23**, 1693– 1710, (August 2020).

[4] Joyjit Chatterjee and Nina Dethlefs, 'A dual transformer model for intelligent decision support for maintenance of wind turbines (to appear)', in *International Joint Conference on Neural Networks (IJCNN)*, Glasgow (UK), (July 2020).

[5] M. Du, S. Ma, and Q. He, 'A scada data based anomaly detection method for wind turbines', in *China International Conference on Electricity Distribution*, Xi'an, China, (August 2016). IEEE.

[6] Guillaume Genthial. Seq2seq with attention and beam search, November 2017.

[7] Raed K Ibrahim, Jannis Tautz-Weinert, and Simon J Watson, 'Neural networks for wind turbine fault detection via current signature analysis', in *WindEurope Summit*, Hamburg, Germany, (September 2016).

[8] Jinhao Lei, Chao Liu, and Dongxiang Jiang, 'Fault diagnosis of wind turbine based on long short-term memory networks', *Renewable Energy*, **133**(C), (10 2018).

[9] Matt MacMahon, Brian Stankieindenergyicz, and Bejamin Kuipers, 'Walk the Talk: Connecting Language Knowledge, and Action in Route Instructions', in *Proc. of National Conference on Artificial Intelligence (AAAI)*, Boston, Massachusetts, (2006).

[10] Wei Qiao and Dingguo Lu, 'A survey on wind turbine condition monitoring and fault diagnosis–part ii: Signals and signal processing methods', *IEEE Transactions on Industrial Electronics*, **62**(10), 6546–6557, (2015).

[11] Somayajulu G. Sripada, Ehud Reiter, Ian Davy, and Kristian Nilssen, 'Lessons from deploying nlg technology for marine weather forecast text generation', in *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, pp. 760–764, (2004).

[12] Yingying Zhao et al., 'Fault prediction and diagnosis of wind turbine generators using scada data', *Energies*, **10**(8), 1210, (2017).

[13] Yajie Yu, Hui Cao, Shang Liu, Shuo Yang, and Ruixian Bai, 'Image-based damage recognition of wind turbine blades', in *2nd International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 161–166, Hefei and Tai'an, China, (August 2017).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 53–54, Santiago de Compostela, Spain, August 29-30, 2020.

OBOE: an Explainable Text Classification Framework

Raúl Antonio del Águila Escobar[1]

Abstract. In this paper, OBOE (explainatiOns Based On concEpts), a framework for explainable text classification is presented. This framework is aimed to provide a process in which explanations, in the form of rules, can be obtained both (a) from the models used to classify the text and (b) from the concepts available in the text and retrieved from the Linked Data Cloud.

1 BACKGROUND AND RESEARCH PROBLEM

1.1 Background

Machine Learning classification algorithms (hereinafter ML algorithms) can be used for knowledge acquisition [18] but not all of them are useful to explain the results obtained, as they are considered 'black boxes' [7].

Black-box models do not provide a clear interpretation of how the model concludes a result [7] but, on the contrary, obtain better evaluation metrics than those obtained by interpretable (e.g. logit) or white-box models (i.e ID3).

Despite performance issues, there are many situations in which it is necessary to provide an explanation of the reasons that lead to a conclusion (such as banking or medical models) [2], even more after regulations such as GDPR [13]. Moreover, in some scenarios, explanations are needed to take actions based on them [7][2].

Defense Advanced Research Projects Agency (DARPA) has founded the eXplainable Artificial Intelligence initiative (XAI), an area which has recently come as one of the main topics of research in the recent years [1] [2][4][7]. Nonetheless, the need to provide meaningful explanations is not new [4][5][6].

There are several views of what does the concept explainability implies and its related concepts such as interpretability [1][7][9]. Also there are new techniques aimed to offer an explainable solution

from any classifier such as LIME [14] or SHAP [11] and others integrated in the classification process itself [10].

Nevertheless, as Lecue points out [8], the research community is far from having a self explainable system which adapts to any data, model, and user or context.

In the Natural Language Processing (NLP) field, the situation is not that different. Liu et al [10] proposed a framework to classify and provide fined grained explanations which can also improve the classifier.

In this sense, Knowledge Graphs (KGs) can play an important role in XAI [8]. Lecue identifies some research areas in which KG might help to overcome some limitations and research challenges. Some of these areas are (i) machine learning and (ii) NLP, by giving semantic layers to classification models and clearing up concepts.

1.2 Research Problem

The aim of the research work in progress is to develop a framework for text classification that eases[2] the creation of further explanations of the results. The main idea is to propose the explanation building based on KGs and Linked Data (LD) paradigm. The research hypothesis (RH) of this work is as follows: *"The identification of the topic of a text can be explained from symbolic rules obtained from the early identification of terms or concepts in the corpus and from the machine learning techniques used to classify them.*

2 PROPOSED SOLUTION

The proposed solution is a framework called OBOE that is based on Positive Unlabeled Learning (PUL) [3] in combination with techniques such as outliers detection [17] and weighted tf-idf [15]. Figure 1 shows the process that is explained in section 2.1:

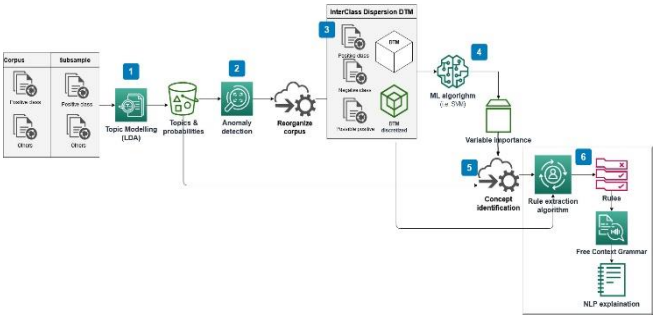

Figure 1. OBOE Framework

2.1 Framework

Research Hypothesis (RH) implies that the system learns to identify texts sharing the same topic by discriminating in first instance words (corresponding to concepts) related to that topic. In this way, terms are relevant to discriminate the topic from the beginning of the process. Positive Unlabeled Learning (PUL) matches this hypothesis. PUL does not require a fully supervised corpus with positive and negative texts. Instead, it uses positive and unlabeled dataset that to early discriminate which texts (and words) do not belong to the positive class (main topic of the text).

[1] Universidad Politécnica de Madrid, email: r.delaguila@alumnos.upm.es

[2] "Eases" means that the different activities of the classification process are oriented to support the explanation of the results obtained.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 55–56, Santiago de Compostela, Spain, August 29-30, 2020.

The PUL strategy used in this framework is the two steps method [3]: the first step is aimed to discriminate negative examples and the second step learns based on both the positive and negative labeled texts identified on the previous step. Instead of using a classic clustering method, the proposed solution uses Latent Dirichlet Allocation (LDA) [16] (step 1), which is based on a similar assumption based on text-topics probabilities distribution.

The proposed solution assumes that there could be texts that would be related in some way to the topic we are trying to discriminate but whose main topic is different (i.e. although the movie review of a film such as 'Gladiator' might be related to Ancient Rome, its main topic is Filmography). Those documents that might be related to the main topic are outliers [17] according to the probability of belonging to positive class (step 2).

The Document Term Matrix (DTM) that will be used by the ML algorithm is based on a tf-idf scheme modified according to an inter class dispersion scheme [15] and discretized in terms of relevance (step 3) for its later use by the Rule Extraction Algorithm. The proposed solution uses as ML algorithms Decision Trees (already used for knowledge acquisition [12]) and Support Vector Machines and we intend to include other models as described in [3]. The ML algorithm is a multi-class classifier with documents belonging to the main topic, documents that might be related to that topic and documents that do not belong to the main topic (step 4).

Finally, the rules are generated by a Rule Extraction Algorithm (step 5). The user can introduce the vocabulary or the number of terms (that will be mapped to concepts of an existing ontology) retrieved by LDA and ML classifier, the length of the rules and the size of the ruleset that that eases to explain the results.

There is also implemented a Free Context Grammar to translate those rules to NLP into Spanish (step 6).

2.2 Current Status and Preliminary Results

Two experiments have been conducted to validate the process. Table 1 details in its columns the variations of the process used in the experiments and the F1 measure obtained.

Table 1. Summary of the experiments.

N.	Corpus(*)		DTM	ML Algotithm	F1
1	Wiki	(i)	TdIdf	CART	69%
2	ARS	(ii)	Weighted Tfidf	SVM	87%

(i) Semiautomatic corpus created from Wikipedia. Positive class: texts belonging to Ancient History (1.159). Negative class: 1.509 texts from other topics. (ii) Amazon Reviews based corpus. Positive class: 50.000 reviews of Books. Negative class: 50.000 reviews of other products.

(*) Texts were chosen randomly for corpus creation.

The rules generated by the Rule Extraction Algorithm are expressed as the following example: `if book = 9 then is a book review with X confidence`, which is translated to `Si la relevancia de book es máxima, es un "book review" con una confianza de X` using the Free Context Grammar.

3 ONGOING WORK AND CHALLENGES

The process is being repeated with two additional corpora and other machine learning algorithms. The terms that are present in the rules are also being mapped to concepts of ontologies. Also, the natural language explanations are translated into English. Some ongoing challenges are (i) to perform an explanation based on similar texts (if there is no rule that can explain the classification of the document), (ii) to perform a further research on the evaluation techniques and methodologies to develop an evaluation module and (iii) to implement an evaluation ontology.

REFERENCES

[1] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[2] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," arXiv:1910.10045 [cs], Dec. 2019.

[3] J. Bekker and J. Davis, "Learning From Positive and Unlabeled Data: A Survey," arXiv:1811.04820 [cs, stat], Nov. 2018.

[4] R. Blanco-Vega, J. Hernández-Orallo, and M. J. Ramírez-Quintana, "El Método Mimético, una Alternativa para la Comprensibilidad de Modelos de 'Caja Negra". *Tendencias de la Minería de Datos en España*. Editorial: Universidad de Sevilla, pp. 391-402 1994.

[5] B. Chandrasekaran and W. Swartout, "Explanations in Knowledge Systems: The Role of Explicit Representation of Design Knowledge," IEEE Expert, vol. 6, pp. 47–49, Jul. 1991, doi: 10.1109/64.87684.

[6] W. Cohen, "Learning Trees and rules with Set-valued Features". Proceedings of AAAI'96, 1996.

[7] D. Gunning, "Explainable Artificial Intelligence (xAI)," Technical Report, Defense Advanced Research Projects Agency (DARPA),2017

[8] F. Lecue, "On The Role of Knowledge Graphs in Explainable AI" Semantic Web Journal, p. 9, 2018.

[9] Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. 61, 2016, doi: 61. 10.1145/3233231.

[10] H. Liu, Q. Yin, and W. Y. Wang, "Towards Explainable NLP: A Generative Explanation Framework for Text Classification," p. 12.

[11] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," presented at the NIPS, 2017.

[12] L. S. Prasanthi and R. K. Kumar, "ID3 and Its Applications in Generation of Decision Trees across Various Domains- Survey," vol. 6, p. 5, 2015.

[13] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJL119/1.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'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 - KDD '16, San Francisco, California, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[15] R.Roul, J. Sahoo, K. Arora." Modified tf-idf term weighting strategies for text categorization" In 2017 14th IEEE India Council International Conference (INDICON), pages 1–6. IEEE, 2017.

[16] R. Rubayyi, A. Khalid, "A survey of Topic Modeling in Text Mining" International Journal of Advanced Computer Science and Applications. 2015, doi: 10.14569/IJACSA.2015.060121.

[17] X. Xu, H. Liu, L. Li, and M. Yao, "A Comparison of Outlier Detection Techniques for High-Dimensional Data," Int. J. Comput. Intell. Syst., vol. 11, pp. 652–662, 2018, doi: 10.2991/ijcis.11.1.50.

[18] K. Zercher and B. Radig, "The Role of Machine Learning in Knowledge Acquisition," Gasteiger J. (eds) Software Development in Chemistry 4. Springer, Berlin, Heidelberg, 1990.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 55–56, Santiago de Compostela, Spain, August 29-30, 2020.

Understandable Deep Learning Analysis for Very High Energy Astroparticle Physics

Ettore Mariotti[1]

Abstract. This work addresses the machine learning problems related to Imaging Air Cherenkov Telescopes analysis with the intent of leveraging a richer description of the captured events by using Convolutional Neural Networks, avoiding the potential information loss due to the hand-crafted feature engineering used in the standard analysis. An explainable and minimal model pointed toward the limitation of the simulation used for training. The full analysis is then tested both on simulated and real data, showing a significant improvement of the sensitivity of the system for different energy bands.

1 INTRODUCTION

In the universe there is more than we can see, our eyes are only sensitive to a small portion of the light spectra but outside of that lie information about a wild variety of phenomena. In particular the MAGIC telescopes[2] are built to capture the most energetics photons (gamma-rays) that reach our planet and interact with the atmosphere. The system is built exploiting a phenomenon known as Cherenkov radiation, a brief and dim blue light emitted by objects traveling faster than the speed of light in the medium they are in. One of the big problems with this technique is being able to discriminate the signal from the background. Unfortunately gamma rays are not the only particles that produce Cherenkov light, as there is a wide and diverse catalog of events that make the telescopes trigger, outnumbering gamma rays at about 2000:1. In order to discover and characterize a new gamma-ray source the analysis must recover only the events caused by gamma photons and then reconstruct their energy and their direction of impact in the sky.

When an energetic particle interact with the atmosphere, it produces a cone of light due to the Cherenkov effect. The telescopes focus with a big array of mirrors this secondary shower of light toward the camera. The camera is composed of 1024 photodetectors which produce a voltage proportional to the light that reach the sensors (raw signal). This voltage is then processed to estimate the number of photons arrived to the camera (Calibrated signal). After a number of triggers that aim to grab plausible gamma events, a set of handcrafted features is computed (Hillas parameters). The physical underlying process is reproduced with a Montecarlo Simulation (MC) which allows to have a set of expected observations with their ground-truth from which the analysis can learn. Once the analysis chain is completed, in order to check its correctness the telescopes are pointed toward the Crab Nebula, a well-known strong and stable source of gamma-rays, and compares its spectra with the one established in the literature by other experiments.

[1] Universitá di Padova, Italy, email: mr.ektor@gmail.com
[2] http://www.magic.iac.es/

In order to select and reconstruct the energy and the direction of the events, the standard analysis makes use of machine learning from a parametrization of the grabbed events. In particular, Aleksic et al. [1] solves the tasks in this way:

1. *Separation*: A random forest for classification is trained on the Hillas parameters for MC gamma rays (Signal) and real observations taken when the telescopes point to a dark region of the sky (Background)
2. *Energy*: A Nearest Neighbour trained on the MC estimates the energy of the events
3. *Direction*: Some additional features are computed with a geometric rationale, and the final result is then fine-tuned with a Random Forest Regressor.

2 DEEP LEARNING ANALYSIS

The feature engineering is (as of any parametrization) destroying information that could be potentially useful for the purposes of the analysis. The approach proposed in this work is based on Deep Learning (DL). More precisely we use Convolutional Neural Network (CNNs) starting from the Calibrated signal. Anyway since the feature engineering provides an already valuable source of predictive information, Hillas parameters are passed to the network as side-information. The architectures for each task thus are composed of two differentiable stages:

1. a CNN body that process the Calibrated signal
2. a Self-Normalizing Neural Network [6] that processes the handcrafted features.

These two branches of processing are then merged together at the final layer in order to produce the final estimation. In particular different CNN bodies were found more suit to different tasks:

1. *Separation* uses a VGG-19 [7]
2. *Energy* uses an Inception V3 [8] with squeeze and excitation blocks [2]
3. *Direction* uses a Densenet-121 [4] with squeeze and excitation blocks [2]

Each model was trained with stochastic weight avearging [5] of checkpointed models computed with a cyclical learning rate [3]. In order to being able to interface with the DL frameworks the calibrated data was interpolated from a native hexagonal grid to a regular rectangular grid. Due to the big memory footprint of the training data ($>$500GB), the events were structured in a custom-designed SQL server.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 57–58, Santiago de Compostela, Spain, August 29-30, 2020.

3 UNDERSTANDING THE SIMULATION/REALITY GAP

When separation is trained and tested on the MC-Real data (as in the standard analysis) the model reach an accuracy of 99.99%, which is unrealistic because we still expect a small 1% of gamma-like events in the Real data.

In order to better investigate what happened in the separation, we designed a minimal and explainable CNN model (called SimplicioNet) based only on the Calibrated signal. Its explainable nature lies in its simple structure:

1. One layer of four convolutional filters 20x20 pixels with ReLU activation.
2. One maxpool layer that selects the most intense activation.
3. One dense layer without bias that combines the values of the previous activation and squeezes the result with a sigmoid.

By having a sigmoid as the last non-linear activation we can know that the feature map combined with a positive coefficient is responsible for the gamma class and the one combined with a negative coefficient is responsible for the non-gamma class. By tracing the location of the most prominent feature (maxpool) it is possible to trace the precise activation field of the Calibrated signal that caused the prediction. When inspected, it was clear that the model was making decisions not by looking at the shower itself, but by looking at other places of the image. The model was solving the separation problem by focusing on some feature of the MC simulation that was different from the Real-Data observations. In other words it was discriminating the Simulation from the Real data instead of gamma from non-gamma. The separation problem was then addressed in two different ways:

- by performing a strong preprocessing of the calibrated signal that cleaned the image;
- by using a MC simulation of the background (which unfortunatly is not really representative of the richness of the real events).

4 RESULTS

The whole pipeline was evaluated both on MC and on Real observations.

Regarding the MC set, we observed that the DL proposal showed significant improvements with respect to the standard analysis. Energy resolution is enhanced of 30% for events above 1 TeV while direction reconstruction is enhanced for each energy band ranging from a 5% to 40%. This is especially important since improvements in the direction reconstruction translates proportionally to improvements in the sensitivity of the system. The whole pipeline, thus including also the separation stage, globally enhance the sensitivity, particularly in the lower energies bands (Fig. 1).

The final pipeline was then evaluated on a set of 5h of real observations of the crab nebula. The results showed performances comparable with the one of the standard analysis, thus proving the proof of concept that this richer analysis can in principle work in a real environment. Most importantly, as the simulation will become more and more realistic it is reasonable to expect a much more dramatic improvement in line with the results on the MC set.

5 CONCLUSIONS

This work leveraged the steep advancements of computer vision by adapting DL solutions to the specific field of Imaging Air Cherenkov Telescopes. Results on MC showed superior performances in terms of sensitivity of the instrument. The implementation of such approach would allow to require less observation time in order to claim a discovery, which translates in more discoveries in the same operational time. Moreover the results point toward improvement in certain energy bands that allows the discovery of previously-undetectable sources. SimplicioNet made evident the limits of the simulation, thus investment of energy and time in its refinement will be a valuable investment as it will resolve in higher performances on the real observations.

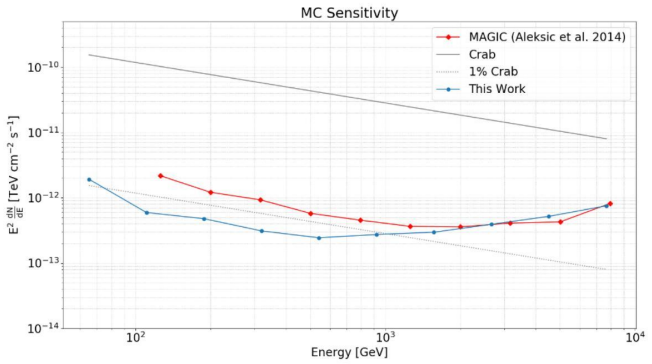

Figure 1. Sensitivity computed on the MC set obtained with the proposed approach (blue line) and compared with the one actually in operation for the telescopes (red line). The lower the better.

ACKNOWLEDGEMENTS

We would like to thank NVIDIA for donating appropriate hardware with which was possible to conduct this investigation. We would also like to thank the software board of the experiment for the valuable feedback obtained at different stages of the project.

This presentation is associated to the NL4XAI project. 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

[1] J. Aleksić, S. Ansoldi, L.A. Antonelli, P. Antoranz, A. Babic, P. Bangale, M. Barceló, J.A. Barrio, J. Becerra González, W. Bednarek, and et al., 'The major upgrade of the magic telescopes, part ii: A performance study using observations of the crab nebula', *Astroparticle Physics*, **72**, 76–94, (Jan 2014).

[2] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2017.

[3] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free, 2017.

[4] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.

[5] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2018.

[6] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks, 2017.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 57–58, Santiago de Compostela, Spain, August 29-30, 2020.

Explaining Bayesian Networks in Natural Language

Conor Hennessy [1]

Abstract. The advent of Artificial Intelligence has been one of the great achievements in Computer Science. The increase in usage of AI systems has given birth to a new challenge; explainability. This extended abstract serves to highlight the importance of explainability in AI, and to examine the challenge it poses. The abstract will have a particular focus on Bayesian Networks and their usage. I will discuss the state of the art in approximate reasoning, namely Bayesian Reasoning and Fuzzy Syllogistic Reasoning, as well as describe a proposed approach to the Natural Language Explanation of Bayesian Networks.

1 INTRODUCTION

In recent times there has been an obvious explosion in the usage of AI models in various situations. In the case of highly complex models, the reasoning or methodology behind the decisions of a model may remain unknown to the end user. It is true that there may exist certain use cases where such a scenario of a lack of understanding about the decision process of a model is acceptable. What is clear, however, is that the potential inability to explain the reasoning behind a model's results serves as a roadblock to the future usage of machine learning in general. In fact, the inability to produce an explanation of the decision of a model may even impinge on the rights of a European citizen in certain circumstances [15] .

A natural example of the conundrum of explainability can be drawn from the medical field. Take for example a model which has learned to diagnose chest pathoses such as lung nodules or cancer, pulmonary embolism or airway diseases [10]. If such a tool were being used by a medical professional to diagnose a patient based on the patient's X-ray, the question of the accuracy would, of course, be a crucial one. Explainability, however, would also be highly important. The medical professional should feel confident in the reasoning of the model and the patient should have peace of mind that the diagnosis is reliable and logical. A method of generating Natural Language explanations of such a model reasoning would alleviate this problem, and is the motivation for this extended abstract.

The model that is the focus of this abstract is the Bayesian Network (BN). In addition to medical applications [7], BN's have found use in evidential reasoning in Law [2] and in DNA profiling in criminal forensics [5], amongst other uses. I will focus on a use case of predictive inference in this abstract. While Using BN's as a model carries certain advantages, which shall be discussed, its use also carries its own challenges. The graphical nature of a BN can aid in the intuition of a system, and there are several graphical tools, such as GeNIe, that can provide understanding for the user. Graphical aids, however, do not provide sufficient knowledge about the model for users not familiar with probabilistic reasoning. Graphs can be misleading, and conditional probability tables are not digestible for an average user. Bayesian reasoning in particular is challenging and often not intuitive.

With an automatic Natural Language Generation method to explain the BN, there can be much more widespread utilization of BN's and higher level of clarity in their usage. I will outline the inner workings of a predictive BN, how the information of a BN can be condensed into Fuzzy quantified syllogisms, and how this could then be used in the Natural Language Explanation.

2 KNOWLEDGE REPRESENTATION AND REASONING

Before an approach to Natural Language Generation can be used to explain a BN, the knowledge and reasoning of a BN must be represented so that a computer system can understand them. Bayesian Reasoning is what is represented in the BN. The starting point for my research will be to model the BN as a quantified syllogism and to solve this syllogism. To do this, a method of representing Bayesian Reasoning must be used. As a first step, I will be investigating Fuzzy Quantified Syllogism.

2.1 Bayesian Reasoning and Bayesian Networks

In classical Bayesian Inference, the Posterior Probability is calculated by including a prior parameter, designed to be a measure of the prior information that is known, which can then be updated as more evidence is gathered [6]. Bayesian networks are Directed Acyclic Graphs (DAG's), representing a set of Random Variables and the dependencies between these Random Variables [11]. The Random Variables are shown as nodes in the DAG, while the edges between these nodes demonstrate the dependencies. The state of a node might affect the probability of another node, depending on the relationship between the nodes. Similarly, the probability that a node is in one state depends on the state of another node according to prior information about the relationships among the nodes. The direction of the arrow on each edge indicates the direction of causality [4], though for certain models, such as those predicting risk, edges may not indicate causality.

Using BN's as a model is advantageous in several ways. They are excellent in modelling potential cause and effect [4] and in risk prediction [1]. They also allow for the encoding of prior knowledge in the model, and the explicit definition of dependencies between nodes leads to more compact graphs [1].

In [1] a predictive BN is introduced, outlining the relationship between patient characteristics, disease and test results. Using the model, predictions can be made for the patient regarding the disease

[1] Centro Singular de Investigacion en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, email: conor.hennessy@usc.es

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 59–60, Santiago de Compostela, Spain, August 29-30, 2020.

in question. It serves as an excellent example for visualising the process, as the predictors, the hierarchy of dependencies and the conditional probability table are clearly presented. It is important to note that the methodology for representing a predictive BN introduced in this abstract is also a potential methodology for diagnostic and inter-casual BN's. The first step I will be investigating is how to represent this information in quantified statements.

2.2 From Bayesian Reasoning to Fuzzy Syllogistic Reasoning

There have been several proposals of methodologies to represent information contained in a BN, such as the qualified support graph method discussed in [3]. The use of Fuzzy syllogisms as a method to represent BN's, however, carries the following potential advantages.

Unlike Boolean Logic, where the output is either true or false, Fuzzy Logic can represent levels of uncertainty or vagueness, as the output can be any $x \in [0, 1]$, where $x \in R$. Natural Language also involves this uncertainty and vagueness. In the case of quantitative linguistic variables, such as with the example of height in [13], they too can be often be fuzzy variables. In the given example of height, tall, very tall, and very very tall all signify varying levels of precision. In the same way, probabilistic statements can be considered fuzzy, for example "approximately .." or "highly unlikely..". As fuzzy logic mimics the human decision making process more closely, it has found use in AI as a form of approximate reasoning [12]. For this reason, it could serve as an excellent tool to bridge the gap between the BN reasoning process and formulating a human language explanation.

Though the connection between probability theory and Fuzzy Logic is a long established one [14], there has been renewed interest in the relationship between Bayesian Reasoning and Fuzzy Syllogistic Reasoning. In order to connect the two concepts, I will take as a starting point the enhanced Syllogistic Reasoning Schema introduced in [8], which allows for any number of variables and several quantifier types, allowing the conditions for the BN to be represented. In [9] a blueprint is then laid out for extracting quantified statements from a BN to form a Fuzzy Syllogism, and compiling a knowledge base for use in an approach to creating a natural language explanation of the BN.

3 EXPLAINABILITY

Once the Bayesian Reasoning process is translated to a Fuzzy Quantified Syllogism, and the Knowledge Base is generated, this would form the first stage in the process of the automatic Natural Language explanation of the BN.

In [2], a 4-step architecture for the Natural Language explanation of a BN is outlined. The Fuzzy Syllogistic approach to BN explanation discussed in [9] focuses on Step 1 from [2], namely the Content Determination phase. Content Determination amounts to the extraction of the information required for the Natural Language Explanation, utilizing the quantified statements.

In the case of template based Natural Language Generation, steps 2 - 4 are normally combined into a single step, namely Linguistic Representation [9]. The Fuzzy Syllogistic approach described above would form the first step in a template based automatic description system, followed by the Linguistic Representation step. In the Linguistic Representation step, a full Natural Language Explanation is generated using the quantified statements compiled in the previous step.

4 FUTURE WORK

As this abstract serves only as a presentation of the problem of explainability of BN's, I will be studying the topics introduced here in more depth. Following this, I hope to define and implement a methodology for explaining in Natural Language the various types of BN reasoning, including predictive, diagnostic and inter-causal reasoning. As BN's grow, scalability becomes an issue, with both the graphical and probability table components of the BN rapidly expanding to cumbersome levels. This issue must also be addressed in future work. The Natural Language Generation methodology can then be implemented in both academic and industrial use cases, with the testing/validation process being assessed by a human user.

ACKNOWLEDGEMENTS

This extended abstract is associated with the NL4XAI project. 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

[1] Paul Arora, Devon Boyne, Justin Slater, Alind Gupta, Darren Brenner, and Marek Druzdzel, 'Bayesian networks for risk prediction using real-world data: A tool for precision medicine', *Value in Health*, **22**, (03 2019).
[2] Jeroen Keppens, 'Explaining bayesian belief revision for legal applications', in *JURIX*, (2016).
[3] Jeroen Keppens, 'Explainable bayesian network query results via natural language generation systems', pp. 42–51, (06 2019).
[4] Lexin Liu, 'A software system for causal reasoning in causal bayesian networks', (2008).
[5] Julia Mortera, Alexander Dawid, and Steffen Lauritzen, 'Probabilistic expert system for dna mixture profiling', *Theoretical population biology*, **63**, 191–205, (06 2003).
[6] Svein Nyberg, *Bayes' Theorem*, 107–135, 08 2018.
[7] Bo Pang, David Zhang, Naimin Li, and Kuanquan Wang, 'Computerized tongue diagnosis based on bayesian networks', *IEEE Transactions on Biomedical Engineering*, **51**, (11 2004).
[8] Martin Pereira, Juan Vidal, F. DÍaz-Hermida, and Alberto Bugarín, 'A fuzzy syllogistic reasoning schema for generalized quantifiers', *Fuzzy Sets and Systems*, **234**, 79–96, (11 2014).
[9] Martín Pereira-Fariña and Alberto Bugarín, 'Content determination for natural language descriptions of predictive bayesian networks', in *11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)*, pp. 784–791. Atlantis Press, (2019/08).
[10] Edwin J. R. van Beek and John T. Murchison, *Artificial Intelligence and Computer-Assisted Evaluation of Chest Pathology*, 145–166, Springer International Publishing, Cham, 2019.
[11] Davy Weissenbacher, 'Bayesian network, a model for nlp?', (01 2006).
[12] John Yen and Reza Langari, *Fuzzy logic: intelligence, control, and information*, volume 1, Prentice Hall Upper Saddle River, NJ, 1999.
[13] L. A. Zadeh, 'Outline of a new approach to the analysis of complex systems and decision processes', *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-3**(1), 28–44, (1973).
[14] Lotfi A. Zadeh, 'Discussion: Probability theory and fuzzy logic are complementary rather than competitive', *Technometrics*, **37**(3), 271–276, (1995).
[15] Scott Zoldi. Explainable ai: Implications for compliance with gdpr and beyond. https://www.linkedin.com/pulse/explainable-ai-implications-compliance-gdpr-beyond-scott-zoldi/. Accessed: 2020-06-20.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 59–60, Santiago de Compostela, Spain, August 29-30, 2020.

Argumentation-based Interactive Factual and Counterfactual Explanation Generation

Ilia Stepin[1]

Abstract. The need for explaining automatically made decisions has shaped a vast body of research on eXplainable Artificial Intelligence (XAI). In addition to stating why a system made a particular ("factual") decision, explaining why a range of alternative non-occurring ("counterfactual") decisions were not made is hypothesized to enhance confidence in the given outcome. Besides, counterfactual explanation suggests the end user a recommendation on how to achieve a favourable result while requiring minimal effort to do so. The present doctoral project approaches the task of factual and counterfactual explanation generation for the multi-label classification problem by discovering the explanatory potential of model's internal structure as well as that of dialogic human-machine interaction on the basis of an argumentation-based framework.

1 INTRODUCTION

Nowadays intelligent systems tend to make a wide use of data-driven algorithms to make automatic decisions. Hence, it is essential that end users understand the underlying reasoning of such systems to be able to trust and effectively manage them [1]. Furthermore, the need to explain automatically made decisions tends to be formalised at the regulatory level worldwide (e.g., in the European Union [6]).

A given automatic decision can be explained factually, i.e., by providing an end user with the data patterns found to be most relevant for a reasoning algorithm when making the decision. However, this approach to explanation is argued to merely summarize the decision-related data instead of really explaining the agent's decision [12]. Alternatively, a complete understanding of why the decision was made may require an explicit specification of why non-made decisions were rejected [10]. Formalised in natural language as conditional sentences, such so-called counterfactual explanations (or counterfactuals, for short) aim at finding a combination of the feature-value pairs minimally different from the actual input, which would enforce the reasoning algorithm to give out a different outcome.

Counterfactuals are known to be context-dependent [9]. Hence, to provide counterfactual explanation is to find the most relevant pieces of counterfactual information and present them to the end user in an effective manner. It is thus of crucial importance to (i) carefully estimate the relevance of the selected counterfactual(-s) to the given automatic decision and (ii) design a human-machine interaction protocol to provide for effective communication of the inferred explanatory pieces of information.

The problem of counterfactual relevance estimation may be viewed from two perspectives. On the one hand, the model's internal structure (if available) can be inspected to assess the proximity of a counterfactual piece of information to the one representing a factual decision. For instance, interpretable classifiers (e.g., decision trees) provide means for assessing the proximity of a counterfactual piece of information to the one representing a factual decision in the context of the model's characteristics. On the other hand, the end user can be placed at the center of the explanation generation problem solving. Such a user-centric approach provides for incorporating user-specific context information (e.g., task-based constraints, user's profile and preferences for specific pieces of counterfactual information).

Paired with a factual explanation, a set of the most relevant counterfactuals is hypothesised to exhibit a transparent big picture of the system's reasoning. A combination of model-specific internal aspects based on user-dependent preferences is expected to enable intelligent systems to produce explanatory output and make a step towards hybrid collective intelligence [5, 11].

The problem of counterfactual explanation generation has recently attracted attention from various research groups. However, some of these approaches do not recognise the problem of counterfactual explanation relevance [7, 15] while others [8, 13] are limited to offering one-shot explanations and do not involve the user explicitly in the process of explanation generation. The present doctoral project aims to overcome this obstacle and synthesize the two aforementioned strategies to address the problem of factual and counterfactual explanation generation.

2 PRELIMINARY RESULTS

At first, a conceptual framework was suggested for generating factual and counterfactual model-specific feature-based explanations for crisp decision trees and fuzzy rule-based classification systems (FRBCS). The method proposed estimates the relevance of counterfactuals on the basis of the minimal XOR-based distance between the test instance vectors and a so-called path-condition matrix containing binarised feature-value permutations. In addition, decision tree-based explanations are linguistically approximated to allow for a consistent pairwise analysis with the equivalent output of a given FRBCS.

The initial experiments have been carried out to generate and evaluate one-shot explanations in the frame of a multi-label classification problem [17]. The results emphasize the importance of measuring counterfactual explanation relevance for rule-based classifiers. Thus, the complexity of the model selected was observed to affect the diversity of candidate counterfactuals as well the compactness of the explanations for the same data instance. In addition, it was concluded that fuzzy classifiers are capable of providing more relevant explanations than their linguistically approximated analogues retrieved from crisp decision trees.

[1] University of Santiago de Compostela (Spain), email: ilia.stepin@usc.es

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 61–62, Santiago de Compostela, Spain, August 29-30, 2020.

3 FUTURE WORK

3.1 A perspective of granular computing

The potential of the suggested method for computing model-based explanations is to be further investigated in the context of granular computing. Information granulation is a form of data abstraction that is argued to be inherent in human cognition [18]. In a general sense, granulation is referred to as "a grouping of elements based on their indistinguishability, similarity, proximity or functionality" [2]. In light of this definition, the potential of the generated explanations is expected to be estimated with respect to their capacity to interpret information granules representing factual and counterfactual information about the classified objects.

A series of experiments are planned to be carried out to assess the potential of the method proposed previously for interpretability of fuzzy information granules. First, the relevance of candidate counterfactuals is to be estimated under the effect of applying a set of discrete α-cut values on the basis of the selected FRBCS given the expert-generated knowledge base. Second, it is proposed to investigate the effect of linguistic approximation of fuzzy sets on the explainability of information granules. Third, the resulting explanations are to be evaluated against those generated given the data-adapted fuzzy sets. The experiments listed above are expected to make a significant contribution to the theory of granular computing in the context of XAI.

3.2 Argumentative counterfactual reasoning

In addition to the core model-based method for generating factual and counterfactual explanations, a cognitively inspired human-in-the-loop architecture is planned to enrich the method proposed above. For instance, a variety of argumentation techniques [3] may serve as a reasoning mechanism over a set of possible counterfactual explanations. Indeed, computational argumentation presents an effective way of logical reasoning as well as a human-machine interface when explaining the output of a multi-agent system (MAS) [14]. Furthermore, recent progress in conversational agent development [4] offers the user the opportunity to affect the quality of output explanations.

Human-machine interaction is claimed to be greatly enhanced through integration of counterfactual explanation generation modules [16]. To enable a MAS with the ability of counterfactual reasoning, a diverse set of possible counterfactual explanations is planned to be regarded as arguments in the predesigned abstract argumentation framework. The candidate counterfactual explanations are thus to be considered claims, argued for or against, to shape the final explanation on the basis of available information about the user as well as their preferences. Achieved as a consensus in a dialogue between the user and the system, a set of factual and the most relevant counterfactual explanations is hypothesised to greatly enhance collective human-machine decision-making. Therefore, further research on methods of user-centric counterfactual explanation generation is believed to be an indispensable step in developing next generation intelligent systems.

To conclude with, it is worth noting that both model-based and argumentation-based methods will be subject to human evaluation. Thus, the argumentative conversational agents are expected to be developed and experimentally validated in real-life use cases in the health care domain.

ACKNOWLEDGEMENTS

This doctoral project is supervised by Dr. Jose M. Alonso and Dr. Alejandro Catala, in collaboration with Dr. Martín Pereira-Fariña. The ongoing research is supported by grant ED431F2018/02 from the Galician Ministry of Education, University and Professional Training and is aligned with the following research projects: EXplica-IA (ED431F2018/02), ADHERE-U (RTI2018-099646-B-I00), and NL4XAI (H2020-MSCA-ITN-2019-860621).

REFERENCES

[1] Amina Adadi and Mohammed Berrada, 'Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)', *IEEE Access*, **6**, 52138–52160, (2018).
[2] Andrzej Bargiela and Witold Pedrycz, *Human-Centric Information Processing Through Granular Modelling*, Studies in Computational Intelligence, Springer Berlin Heidelberg, 2009.
[3] Günther Charwat, Wolfgang Dvořák, Sarah A. Gaggl, Johannes P. Wallner, and Stefan Woltran, 'Methods for solving reasoning problems in abstract argumentation – a survey', *Artificial intelligence*, **220**, 28–63, (2015).
[4] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang, 'A survey on dialogue systems: Recent advances and new frontiers', *ACM SIGKDD Explorations Newsletter*, **19**(2), 25–35, (2017).
[5] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister, 'Hybrid intelligence', *Business & Information Systems Engineering*, **61**(5), 637–643, (2019).
[6] European Commission, 'Artificial Intelligence for Europe', Technical report, European Commission, (2018). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions.
[7] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, 'Factual and Counterfactual Explanations for Black Box Decision Making', *IEEE Intelligent Systems*, **34**(6), 14–23, (2019).
[8] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata, 'Grounding visual explanations', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–286. Springer, (2018).
[9] Jonathan Ichikawa, 'Quantifiers, knowledge, and counterfactuals', *Philosophy and Phenomenological Research*, **82**(2), 287–313, (2011).
[10] Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, **267**, 1–38, (2019).
[11] Marieke M.M. Peeters, Jurriaan van Diggelen, Karel van den Bosch, Adelbert Bronkhorst, Mark A. Neerincx, Jan Maarten Schraagen, and Stephan Raaijmakers, 'Hybrid Collective Intelligence in a Human–AI Society', *AI & Society*, 1–22, (2020).
[12] Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, **1**(5), 206–215, (2019).
[13] Chris Russell, 'Efficient Search for Diverse Coherent Explanations', in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, pp. 20–28, (2019).
[14] Elizabeth I. Sklar and Mohammad Q. Azhar, 'Explanation through argumentation', in *Proceedings of the 6th International Conference on Human-Agent Interaction*, pp. 277–285. Association for Computing Machinery, (2018).
[15] K. Sokol and P. Flach, 'Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant', in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5868–5870, (2018).
[16] Kacper Sokol and Peter Flach, 'One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency', *KI - Künstliche Intelligenz*, 235–250, (2020).
[17] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martin Pereira, 'Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers', in *IEEE World Congress on Computational Intelligence*, pp. 1–8, (2020).
[18] L Zadeh, 'Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic', *Fuzzy Sets and Systems*, **90**(2), 111–127, (1997).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 61–62, Santiago de Compostela, Spain, August 29-30, 2020.

Age-Related Individual Differences and their Implications for the Design of Cognitive Interactive Systems

Lucas Morillo-Méndez[1]

Abstract. In the fourth industrial revolution we are witnessing the appearance of new interactive technologies designed to support older adults (OA). This technology has mostly focused on assistance based on age-derived issues and has put a special emphasis on offering solutions to them. The goal of this project is to investigate how age-related differences affect the interaction between a human and the system —be it a robot or a virtual agent— in terms such as engagement with the interaction, processing of social cues and the extent to which a performed task is actually facilitated by the agent. The project focuses on studying the means of interaction with assistive systems between different groups of users —younger and older adults—rather than exploring final applications of technology as a tailored solution for different people.

1 INTRODUCTION

Older Adults (OA) are one of the main groups of users for whom new technologies may offer more opportunities as a consequence of population ageing occurring in parallel to the fourth industrial revolution. By the year 2050, it is estimated that population over 65 years old will represent a 16% of the population [12]. Despite the potential of these technologies, they are mostly oriented to compensate for declines associated with ageing. While these assistive technologies have an outstanding potential, there is also a need to think of OA as a group of users who are not always fragile in need of help. A healthy OA could benefit from the use of technology that could assist in everyday tasks in a similar way than any other population could do. Nevertheless, there are changes at the cognitive level through human lifespan that could modulate the preferences of interaction with these systems. In addition, these changes can create access barriers in the interaction with them. While complex interfaces could heighten the ease of use for OA, multimodal and natural interactions are perceived as more intuitive for everybody [10]. In addition, the use of non-verbal cues can enhance the feeling of engagement with a robot and a virtual agent, and the interaction can be facilitated if these are designed considering human aspects of social cognition and communication [1]. However, there is little theory-driven research that has searched for principles of interaction between cognitive interactive systems and OA based on these traits.

Traditionally, research in the field of human-computer interaction (HCI) and human-robot interaction (HRI) involving OA has put special emphasis on the validation of systems created for specific functions. This approach is useful for exploring the acceptability and usability of a system, but it does not emphasise the creation of human-based models of interaction with technology. These models could be grounded on theory-driven studies of HCI/HRI, but control groups are mostly lacking in the literature [13], making developmental models of interaction with cognitive systems a barely explored possibility. This project will evolve over a series of experiments involving OA and younger adults (YA) with the aim of exploring links between ageing and different aspects of the interaction with cognitive systems, such as how are they perceived, their acceptability and their actual effectiveness. To do so, I will rely on a multidisciplinary approach to the study of HCI/HRI that uses tools and models from the fields of experimental and cognitive psychology to develop a framework to study how age affects the interaction with cognitive systems.

[1] Örebro University, Sweden, email: lucas.morillo@oru.se

2 PREVIOUS STUDIES

There is evidence of users' preference for cognitive systems that show certain human features. It has been shown that anthropomorphic robots enhance the social presence of a system [11], a concept related with the degree of awareness somebody attributes to a partner —or to a system in this case— [9]. The attribution of social presence over a system that features a human-like shape and behaviour is mediated by a top-down attentional bias towards social stimuli in humans that varies in normal ageing [6]. However, this variation have not been tested in interactions with systems featuring social features. Despite our tendency towards social stimuli, just a few studies in HCI/HRI explore aspects of social cognition in humans and the possible differential outcomes in OA [5, 7].

Feingold-Polak et al. [7] investigated the effects of embodied anthropomorphic robots with two groups of OA and YA. Their study showed that despite the preference for humanoid robots in both groups, OA preferred to interact by touching the hand of the robot rather than using a tablet attached to the system, which was the preferred way of interaction for the younger group.

The effects of robots that feature non-verbal communication cues have also been studied and these have generally scored high in acceptability and social presence. Crompton and MacPherson [4] reported an increase in accuracy and decrease in completion time for a collaborative task with a non-embodied agent when it used a natural human voice and the OA thought of the system as a human person. However, we could not find more studies that have focused in non-verbal interaction communication between social robots and OA. Nevertheless, it has been shown that robots that use personality matching through gaze, a non-verbal social cue, increase the motivation of the user to engage in a repetitive task [1]. In spite of not studying interactions with OA, this experiment explore interactions

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 63–64, Santiago de Compostela, Spain, August 29-30, 2020.

with social robots based on individual differences of personality traits (introversion-extroversion, from the Big-Five model of personality [8]), which has been shown to vary through the human lifespan towards lower extraversion attributes during adulthood [3]. This highlight the importance of non-verbal communication for designing better interactions with cognitive systems. It is also an example of how human-human interaction principles (Similarity-Atraction principle [2]) and psychological models [8] can be applied to HCI and HRI.

3 METHODS

A set of experiment will explore if parameters such as deictic eye gaze from the robot (i.e. eye behaviour signalling to a point in space) affect the interaction with the system in a different way based on age. Other variables of interest are embodiment and anthropomorphism, as they constitute social stimuli that could be attended differently as a result of ageing. These variables will be studied by means of interaction with robots, agents within Virtual Reality (VR) and screen-based agents.

This research is based on a more naturalistic approach towards experimental psychology, similar to other experiments in HRI, but with more control over variables and the environment. Eye-tracking will be used in order to quantify features receiving more overt visual attention by participants. Data from questionnaires measuring mental workload, social presence and scales on perception of the agent will be gathered, as well as qualitative data in the form of annotations from comments and behaviours from the participants.

4 EXPECTED RESULTS

This research will help to shape the preferred and optimal means of interaction between cognitive systems —i.e. virtual agents and robots— and OA. Changes in the way OA interact with the systems are expected based on studies from human-human interaction, which would constitute the first steps towards the creation of AI models capable of adapting to the possible limitations of their human partner. In more concrete terms referred to our first experiment, it is expected that deictic eye-gaze behaviour from a robot to be less useful for an OA in terms of aid for task completion or for communicative purposes when compared with a YA.

5 DISCUSSION AND CONCLUSION

This project will delve into how age-related differences shape interaction with cognitive agents. Research in the field of HRI has focused in exploring how OA interact with robots and virtual agents, but this is normally in the interest of validating specific goal-oriented systems. In addition, the general lack of control groups limit causal link between the age of users and the output. In this research, differences between two groups of OA and YA will be addressed when interacting with a virtual agent and a robot. If found, these differences will have implications in the way in intelligent systems are designed. On the one hand, this knowledge might enrich models of adaptive AI if designed for everybody. On the other hand, if designed for OA, this research might inform of which elements are important based on how these are attended and evaluated.

This research aims to reach a wider generalisation of its results by not validating a system based on user's opinion of its usefulness for helping with a specific task, but it is also true that the system cannot be purposeless during the interaction. Consequently, it will be difficult that some of evaluations made by participants are not influenced by the task being performed, which is the main limitation of this project —e.g. even if deictic eye behaviour helps the user in finding an ingredient while preparing a meal, the user may still evaluate the system negatively if the behaviour is not found necessary for performing such a task.

To conclude, it could be argued that if humans are biased towards social stimuli in general, designing a robot that completely emulates humans would allow us to skip the steps of investigating which are the important human-like elements for a robot. This could be true, but also not feasible at this stage. In addition, a more human-like agent might not always be optimal, as it has been shown that for certain tasks, task efficiency might be reduced based on the social features of a system [11].

ACKNOWLEDGEMENTS

I would like to thank my Phd advisor, Oscar M. Mozos, for his guidance during the preparation of this abstract and STAIRS paper [2].

REFERENCES

[1] Sean Andrist, Bilge Mutlu, and Adriana Tapus, 'Look like me: Matching robot personality via gaze to increase motivation', in *Proc. CHI.*, volume 2015-April, pp. 3603–3612. Association for Computing Machinery, (apr 2015).

[2] Virginia Blankenship, Steven M. Hnat, Thomas G. Hess, and Donald R. Brown, 'Reciprocal interaction and similarity of personality attributes', *J. Soc. Pers. Relatsh.*, **1**(4), 415–432, (1984).

[3] Paul T Costa, Jeffrey H Herbst, Robert R McCrae, and Ilene C Siegler, 'Personality at midlife: Stability, intrinsic maturation, and response to life events', *Assessment*, **7**(4), 365–378, (2000).

[4] Catherine J. Crompton and Sarah E. MacPherson, 'Human agency beliefs affect older adults' interaction behaviours and task performance when learning with computerised partners', *Computers in Human Behavior*, **101**, 60–67, (2019).

[5] Juan Fasola and Maja J. Mataric, 'Evaluation of a spatial language interpretation framework for natural human-robot interaction with older adults', in *Proc. - IEEE ROMAN*, pp. 301–308. Institute of Electrical and Electronics Engineers Inc., (nov 2015).

[6] Francesca Federico, Andrea Marotta, Margherita Orsolini, and Maria Casagrande, 'Aging in cognitive control of social processing: evidence from the attention network test', *Aging Neuropsychol. Cogn.*, (2020).

[7] Ronit Feingold-Polak, Avital Elishay, Yonat Shahar, Maayan Stein, Yael Edan, and Shelly Levy-Tzedek, 'Differences between young and old users when interacting with a humanoid robot: A qualitative usability study', *Paladyn*, **9**(1), 183–192, (feb 2018).

[8] Lewis R. Goldberg, 'A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models.', *Pers. Psy. Eur.*, **7**, 7–28, (1999).

[9] Chad Harms and Frank Biocca, 'Internal Consistency and Reliability of the Networked Minds Measure of Social Presence', *7th Int. Wksh. Presence*, 246–251, (2004).

[10] Bran Knowles, Yvonne Rogers, Jenny Waycott, Vicki L. Hanson, Anne Marie Piper, and Nigel Davies, 'HCI and aging: Beyond accessibility', in *Proc. CHI.*, (may 2019).

[11] Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson, 'The effects of anthropomorphism and non-verbal social behaviour in virtual assistants', in *Proc. ACM IVA*, pp. 133–140, (jul 2019).

[12] United Nations, 'World Population Prospects 2019: Highlights', Technical Report 141, (2019).

[13] Oded Zafrani and Galit Nimrod, 'Towards a Holistic Approach to Studying Human-Robot Interaction in Later Life', *Gerontologist*, **59**(1), E26–E36, (2019).

[2] Lucas Morillo-Mendez and Oscar Martinez Mozos 'Towards Human-Based Models of Behaviour in Social Robots: Exploring Age-Related Differences in the Processing of Gaze Cues in Human-Robot Interaction', 9th European Starting AI Researchers' Symposium (STAIRS), in Press, (2020).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 63–64, Santiago de Compostela, Spain, August 29-30, 2020.

Big Data in IoE: investigating IT approaches to Big Data in healthcare whilst ensuring the competing interests of the right to health and the right to privacy

Aiste Gerybaite[1]

Abstract. Big Data has been vastly applied in sports domain while Big Data in IoE in healthcare has only recently started gaining traction. Sophisticated IoE devices are now able to process Big Data in order to monitor undesirable events through real-time alerting such as crash of vital signs with wearable wireless sensors, domestic accidents involving elderly people, or assist in predicting and monitoring outbreaks of various diseases which may put at risk public health as a whole. Due to the multidimensional nature of healthcare emergencies, the use of Big Data in such emergencies poses a number questions, not only with respect to the precise definition of an emergency, the agencies involved, the procedures used, but also questions with respect to securing privacy and the right to health in such emergencies without hindering the potential benefits of the development of Big Data solutions within healthcare sector.

1 PROJECT DESCRIPTION

1.1 Research problem

The correlation between Big Data and data protection is undeniable. Big Data in healthcare brings about several data protection and privacy challenges. Firstly, many of the classic fair information principles and data protection principles are being challenged by Big Data. In fact, van Der Sloot argues that the current data protection regime, based on the principle of data minimization, undermines Big Data's potential as in Big Data scenarios the objective is to collect as much data as possible and store data for as long as possible (for example, in healthcare research Big Data analytics are used to find trends and treatments that have the highest rate of success in the real world for cancer treatment). Further, Big Data tests the principles of *ratione personae* and *ratione materiae*[1]. In the first instance, Big Data do not focus on specific individuals but rather on groups of individuals/everyone, while in the latter, it becomes unclear whether a particular right (right to data protection) is even involved in certain Big Data scenarios. This raises a question on whether data protection and privacy in Big Data scenarios should take into account the collective dimension of data protection rights rather than focus on the individual dimension of such rights[2]. Additionally, since Big Data processing is often transnational, the rapidly changing technology may circumvent the applicable data protection provisions. Consequently, many questions arise whether the definitions of personal data, anonymized data, sensitive personal data are still tenable in the age of Big Data where ICT approaches tend to be more and more dynamic while the law remains faithful to its static nature.

In healthcare sector, the research to date focuses on the development of the ICT solutions for the sector (various medical devices, tracing devices) or on the regulatory requirements applicable to the sector. Yet, Big Data, privacy and data protection issues tend to be overlooked. In particular, the healthcare sector tends to overlook the two dimensions of healthcare emergencies, the public health and the individual dimension, and their implications to data protection and privacy. The public dimension of healthcare emergencies refers to emergencies such as global outbreaks of diseases (such as Covid-19, SARS etc). The latter, instead, refers to loss of vital signs by an individual which would not qualify as a public health emergency but, nevertheless, could have devastating impact on an individual's wellbeing.

The compendium of data aggregated by various actors in such emergencies though the use various health-monitoring, medical devices, or data generated and captured through various tracing apps and other healthcare software systems proliferates both concerns and opportunities. Indeed, a number of research focuses on developing WBAN, Primary Mobile Device (PMD), and Internet using Bluetooth, ZigBee IEEE 802.15.4, or 3G/LTE/Wi-Fi communication technologies-based ICT solutions[3]. Nonetheless, integrating real-time and similar health monitoring data has significant ethical[4], legal and social implications on individual's informational privacy due to the abstract and value based nature of law which is built on compromise. Instead, "Big Data is empirical, algorithmic, and deterministic. Also, Big Data is inherently acontextual. Big Data cannot interpret itself, nor can it discern the indeterminate boundaries of legal principles[5]".

Since the use of Big Data depends on the context it is used in, from the perspective of healthcare Big Data Shameer argues that the ownership rights of the data subject and the control of how patients may opt-in and opt-out are important for building real-time and similar health monitoring systems[6]. Such systems would need to be secure in order to ensure high degree of protection of personal and sensitive data in order to adhere to data protection requirements. This is particularly pertinent in the European scenario where the General Data Protection Regulation (GDPR) is applicable. In this

[1] CIRSFID, University of Bologna/ Law Department of the University of Torino/ Department of Computer Science of the University of Luxembourg; email: aiste.gerybaite@unito.it

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 65–66, Santiago de Compostela, Spain, August 29-30, 2020.

regard, an in-depth analysis of the scope and application of articles 5,6 and 9 of the GDPR is required.

Likewise, ransomware and hacking attacks do not only work on data that a healthcare organisation already holds but also when various devices and health apps transmit, collect and share the data between one another or with a healthcare organisation. Thus, the processing of such personal and sensitive data may lead to data privacy violations. According to Cohen privacy violations can be divided into two categories, consequentialist and deontological concerns[7]. Whilst the first point concerns the negative consequences that the data breach carries with it, the second one does not depend on experiencing the negative consequences (ie. publication of sensitive information on an open web, which inadvertently allows third parties to gain access to such data).Whilst most research encountered addresses the potential breach in terms data privacy under the existing regulatory regime, few assess to what extent the existing ICT solutions in healthcare already provide for data privacy and data security.

In conclusion, the current state of the art undoubtedly indicates a gap between the appropriate translation and application of fundamental legal research into concrete scenarios with specific ICT technologies used in healthcare sector. As the ethical-legal research world focuses on fostering a high-level discussion on Big Data, healthcare and IoE, the ICT sector wonders what all this means to their specific scenario. Finally, the researcher believes that through the proper analysis of case-specific scenarios a further path for the regulation of Big Data and data protection may be paved.

2 RESEARCH OBJECTIVES

The research project investigates ICT approaches to the processing of Big Data in healthcare emergencies focusing on the two dimensions of such emergencies. In particular, the research focuses on applications used to assist and manage pandemics, and on wearables used for tracking person's vital signs and how data is processed and treated in healthcare emergencies. Such analysis permits to provide a multidimensional look at the core issues from both legal and technical standpoints for Big Data in healthcare emergencies with possible solutions for the issues. The objective therefore is to set ground for future regulatory regime which would facilitate the sharing of personal and personal sensitive data in a secure manner that would allow to implement ICT solutions within healthcare sector. Finally, the researcher believes that certain existing ICT tools used in healthcare may already be capable of guaranteeing data protection and privacy.

The research project will engage a discussion on the two angles of Big Data, the regulatory side and the technical side:

1. on the regulatory side the project analyses what part data protection legislation plays within the development of Big Data applications in healthcare sector. The research further looks into the precise notions of an emergency from both empirical and legal standpoints; the notions of data privacy vs data security; and securing privacy and the right to health in such emergencies without hindering the potential benefits of the development of Big Data solutions within healthcare sector.
2. The research considers to what extent the current design guidelines for wearables and various applications in healthcare take into account the applicable data protection framework.
3. The research explores how to balance the competing rights and interests of the affected parties with respect to data privacy and the right to health. In particular, through looking at the specific Big Data applications within healthcare, the research attempts to answer the question of how the benefits of Big Data in healthcare can be captured in a way that respects fundamental principles of ethics and the rights to privacy and health.

In essence, the approach taken within research aims to look at the Big Data in healthcare and regulations with a novel perspective, ie. evaluate specific ICT solutions within the specific legislative framework. If the research is successful, the proposed research would benefit a wide range of stakeholders including legislators, scholars, practitioners, and consumers. The project is aimed and bridging the gap between the law and the ICT solutions within healthcare sector with respect to the EU applicable regulations and the ICT tools developed within the healthcare sector.

3 AUTHOR

Aiste Gerybaite is a doctoral researcher in the Law, Science and Technology Joint Doctorate - Rights of Internet of Everything, funded by Marie Skłodowska-Curie Actions. Her doctoral research focuses on Big Data for Health in IoE in emergency situations from the perspective of the competing interests of the right to health and the right to privacy.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie ITN EJD grant agreement No 814177.

REFERENCES

[1] B. Van der Sloot and A. De Groot, *The handbook of privacy studies: An interdisciplinary introduction*. Amsterdam University Press, 2018.

[2] L. Floridi, 'Open Data, Data Protection, and Group Privacy', *Philos. Technol.*, vol. 27, no. 1, pp. 1–3, Feb. 2014.

[3] M. M. Rathore, A. Ahmad, A. Paul, J. Wan, and D. Zhang, 'Real-time Medical Emergency Response System: Exploiting IoT and Big Data for Public Health', *J. Med. Syst.*, vol. 40, no. 12, p. 283, Dec. 2016.

[4] L. Floridi *et al.*, 'Key Ethical Challenges in the European Medical Information Framework', *Minds Mach.*, vol. 29, no. 3, pp. 355–371, Sep. 2019.

[5] B. van der Sloot and S. van Schendel, 'Ten Questions for Future Regulation of Big Data: A Comparative and Empirical Legal Study', *J. Intellect. Prop. Inf. Technol. Electron. Commer. Law*, vol. 7, 2016.

[6] K. Shameer, M. A. Badgeley, R. Miotto, B. S. Glicksberg, J. W. Morgan, and J. T. Dudley, 'Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams', *Brief. Bioinform.*, vol. 18, no. 1, pp. 105–124, Jan. 2017.

[7] W. N. Price and I. G. Cohen, 'Privacy in the age of medical big data', *Nat. Med.*, vol. 25, no. 1, pp. 37–43, Jan. 2019.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 65–66, Santiago de Compostela, Spain, August 29-30, 2020.

Deep learning neural networks to detect anomalies in video sequences

Jorge García-González[1]

Abstract. The purpose of this work is to summarize a PhD work focused on anomaly detection in video sequences. The first section presents the problem, the second one summarizes the state of the art, the third section exposes our work within the area so far and the fourth one explains our current and future work.

1 PROBLEM

Multimedia data are a sign of our times. Our society generates each day more and more data, including a considerable share of images and videos with very different sources and objectives: smartphones, surveillance and traffic cameras, satellites, handmade or digital art, films, TV shows, X-ray or CT-scan medical images, etc. The increasing amount of image and video data imposes the need for automatic analysis methods and computers barely understand them.

One of the approaches to better understand them and reduce the amount of data that requires human supervision is the detection of elements that should be subject to further study. These elements are what we often call anomalies or outliers.

How do we define an anomaly? That is a key question with an ambiguous answer: it depends on the context. Since we are interested in video sequences, we will focus on their usual context.

Anomaly detection in images and video sequences can lead to two problems. *Foreground Segmentation* problem (also called *Background Subtraction* in bibliography) is the problem of, given a video, classify each pixel in each frame as foreground (anomaly) or background (normal). On a road, each car would be classified as foreground while a tree next to the road would be considered as background even if its branches are also moving with the wind.

Behavior anomaly detection in a video sequence would be the problem of, given a video sequence, classify each pixel in each frame as usual behavior or anomaly. In our previous road example, if usually the cars are going from left to right, then a car moving from right to left would be an anomaly. If there are no motorcycles but one, then that one could also be considered as an anomaly. This second problem implies a deeper understanding of the video and is more challenging.

2 STATE OF THE ART

Foreground segmentation algorithms have been proposed for decades. First methods were based on pixel level analysis [14] but they show very low noise robustness so other later methods also classify at pixel level but use information from surrounding pixels to increase noise robustness [12, 9].

[1] University of Málaga, Spain, email: jorgegarcia@lcc.uma.es

In recent years deep learning advances have provided other approaches to *foreground segmentation* based on a semantic segmentation network output [1] or by using autoencoders networks to increase noise robustness with no dependence of a network trained in a supervised way [17].

Behavior anomaly detection methods often study characteristics as Histogram of Optical Flow [2], use Markovian techniques [8] or are based on statistics aggregates [11]. Other newer methods use deep learning [16, 18] and it is common to use autoencoders networks [15, 13] in a way pretty similar to *foreground segmentation* methods. This common approach is based on the reduction of data dimensionality in order to analyze that lower dimension data. It is important to note that the autoencoder approach to both problems usually uses only autoencoder's encoding layers, so it acts as a *Principal Component Analysis* (PCA) method. Figure 1 shows the usual basic scheme for this approach. This approach shows the following advantages: Autoencoders are trained in an unsupervised way, so no labeled data is needed; by deleting non-relevant data, the noise robustness is increased; the amount of data to analyze is much more manageable.

3 OUR WORK

Following previous works in the field as [17], our main approach to the *foreground segmentation* problem is based on the use of deep learning techniques to change the data representation dimension into a lower one in combination with other traditional analysis methods often based on probability theory or statistical analysis. As previously mentioned, the lower dimensionality representation has the advantage of the destruction of the smallest details since they usually have no relevance to the anomaly detection in the image but are a source of noise.

In our preliminary work [5], we studied how a simple statistical analysis over a lower dimension representation based on autoencoders could lead to a method able to deal with Gaussian noise in the video sequence. Following that path, on [6] we improved the codification output analysis by using a Gaussian model, and we tested the newly proposed method against different kinds of noises, including Gaussian, uniform and salt and pepper. However, this approach has a problem: since we need to split the images into patches in order to encode them, the output will have a low segmentation resolution.

To avoid that problem, in [3], we proposed a new method using shifted tilings to increase both robustness and resolution. In [4], we include the tiling approach in combination with a new Gaussian analysis, so our method is robust but also has higher segmentation resolution than [5] and [6]. Figure 3 shows a comparison between two *foreground segmentation* methods with different robustness to noise.

In parallel, we have tried other different approaches to *foreground*

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 67–68, Santiago de Compostela, Spain, August 29-30, 2020.

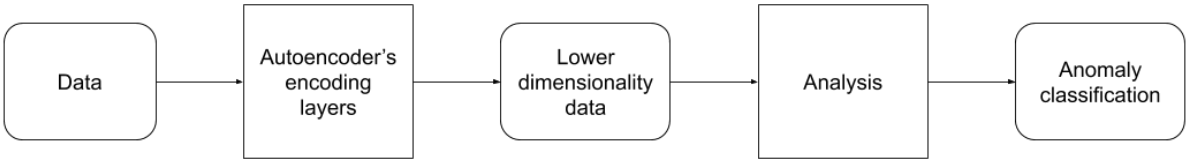

Figure 1. Basic usual autoencoder approach used both in *foreground segmentation* and *behavior anomaly detection* in video sequences.

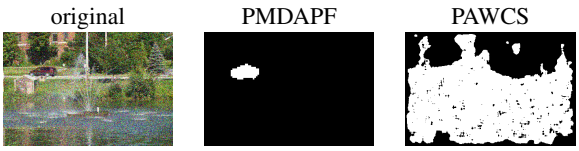

Figure 2. *Foreground segmentation* example. On the left, the original image from changedetection.net dataset with added noise, on the center, our PMDAPF [4] method output, on the right, PAWCS [12] method output.

segmentation. In [10], we studied how some simple preprocessing as downsampling the images in order to apply simple foreground detection methods could also lead to robust results. Another parallel approach will be presented at the ECAI conference. In that work, we show a way to combine the information from a semantic segmentation network as Mask-RCNN [7] with a probabilistic model to obtain *foreground segmentation*.

4 FUTURE WORK

We are currently working on methods to deal with *behavior anomaly detection* based on the approach we have followed in dealing with *foreground segmentation*. As we noted before, the problem is much harder since, in *foreground segmentation*, the background (context) does not change too much, and even if it contains movement, it can be countered with noise robustness. *Behavior anomaly detection* has a context much more dynamic. Spatio-temporal information is much more important, and so the way to model the context is much more challenging, but we think that it also has more applications. We are also currently developing another *foreground segmentation* method that contains most of the advantages from [4] but with no resolution limit.

ACKNOWLEDGEMENTS

The author's thesis advisors are Juan Miguel Ortiz de Lazcano Lobato and Rafael Marcos Luque Baena from University of Málaga. The following work was accepted in ECAI conference:

Jorge García-González, Juan M. Ortiz-De-Lazcano-Lobato, Rafael M. Luque-Baena and Ezequiel López-Rubio, 'Foreground detection by probabilistic mixture models using semantic information from deep networks', 24th European Conference on Artificial Intelligence (2020).

REFERENCES

[1] M. Braham, S. Piérard, and M. Van Droogenbroeck, 'Semantic background subtraction', in *IEEE International Conference on Image Processing (ICIP)*, pp. 4552–4556, Beijing, China, (September 2017).

[2] Yang Cong, Junsong Yuan, and Ji Liu, 'Sparse reconstruction cost for abnormal event detection', *CVPR 2011*, 3449–3456, (2011).

[3] Jorge García-González, Juan M. Ortiz-De-Lazcano-Lobato, Rafael Marcos Luque Baena, and Ezequiel López-Rubio, *Background Modeling by Shifted Tilings of Stacked Denoising Autoencoders*, 307–316, 05 2019.

[4] Jorge García-González, Juan M. Ortiz-De-Lazcano-Lobato, Rafael Marcos Luque Baena, and Ezequiel López-Rubio, 'Background subtraction by probabilistic modeling of patch features learned by deep autoencoders', *Integrated Computer-Aided Engineering*, 1–13, (03 2020).

[5] Jorge García-González, Juan M. Ortiz-De-Lazcano-Lobato, Rafael Marcos Luque Baena, Miguel A. Molina-Cabello, and Ezequiel López-Rubio, *Background Modeling for Video Sequences by Stacked Denoising Autoencoders: 18th Conference of the Spanish Association for Artificial Intelligence, Proceedings*, 341–350, 01 2018.

[6] Jorge García-González, Juan M. Ortiz-De-Lazcano-Lobato, Rafael Marcos Luque Baena, Miguel A. Molina-Cabello, and Ezequiel López-Rubio, 'Foreground detection by probabilistic modeling of the features discovered by stacked denoising autoencoders in noisy video sequences', *Pattern Recognition Letters*, **125**, 481–487, (07 2019).

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, 'Mask r-cnn', in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, (Oct 2017).

[8] Jaechul Kim and Kristen Grauman, 'Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates', *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2928, (2009).

[9] E. López-Rubio, R.M. Luque-Baena, and E. Domínguez, 'Foreground detection in video sequences with probabilistic self-organizing maps', *International Journal of Neural Systems*, **21**(3), 225–246, (2011).

[10] Miguel A. Molina-Cabello, Jorge García-González, Rafael Marcos Luque Baena, and Ezequiel López-Rubio, 'The effect of downsampling–upsampling strategy on foreground detection algorithms', *Artificial Intelligence Review*, (02 2020).

[11] Venkatesh Saligrama and Zhu Chen, 'Video anomaly detection based on local statistical aggregates', *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2112–2119, (2012).

[12] P. St-Charles, G. Bilodeau, and R. Bergevin, 'Universal background subtraction using word consensus models', *IEEE Transactions on Image Processing*, **25**(10), 4768–4781, (2016).

[13] Jun Wang and Limin Xia, 'Abnormal behavior detection in videos using deep learning', *Cluster Computing*, **22**(4), 9229–9239, (Jul 2019).

[14] C.R. Wren, A. Azarbayejani, T. Darrell, and A. Pentl, 'Pfinder: Real-time tracking of the human body', *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **19**(7), 780–785, (1997).

[15] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe, 'Learning deep representations of appearance and motion for anomalous event detection', in *Proceedings of the British Machine Vision Conference (BMVC)*, eds., Mark W. Jones Xianghua Xie and Gary K. L. Tam, pp. 8.1–8.12. BMVA Press, (September 2015).

[16] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe, 'Detecting anomalous events in videos by learning deep representations of appearance and motion', *Computer Vision and Image Understanding*, **156**, 117–127, (2017).

[17] Yaqing Zhang, Xi Li, Zhongfei Zhang, Fei Wu, and Liming Zhao, 'Deep learning driven blockwise moving object detection with binary scene modeling', *Neurocomputing*, **168**, (06 2015).

[18] Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, and Zhijiang Zhang, 'Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes', *Sig. Proc.: Image Comm.*, **47**, 358–368, (2016).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 67–68, Santiago de Compostela, Spain, August 29-30, 2020.

Embedding based Link Prediction for Knowledge Graph Completion

Russa Biswas[1]

Abstract. This thesis proposes a novel Knowledge Graph (KG) embedding model for Link Prediction (LP) for Knowledge Graph Completion (KGC). The missing links in a KG are predicted based on the existing contextual information as well as textual entity descriptions. The model outperforms the state-of-the-art (SOTA) model DKRL for FB15k and FB15k-237 datasets.

1 INTRODUCTION

KGs are large networks of real world entities and relationships between them. The facts are represented as a triple $< h, r, t >$, where h and t are the head and tail entities respectively and r represents the relation between them. Despite the huge amounts of relational data, KGs are sparse and often incomplete as the links between the entities are missing. Furthermore, different KGs have information about the same real world entities but the fact that these entities in different KGs are same is missing.

LP is a fundamental task of KGC that aims to estimate the likelihood of the existence of links between entities based on the current observed structure of the KG. LP task can be performed across different KGs to predict the missing links between two same entities across KGs and is also known as Entity Alignment. This thesis focuses on the KGC task based on predicting the missing links within the KG as well as across multiple KGs.

Due to the high complexity of the graph mining algorithms, the latent representation of a KG is learned into a low dimensional space. To-date many algorithms are proposed to learn the embeddings of the entities and relations into the same vector space. However, none of the SOTA models consider the contextual information of the KGs along with the textual entity descriptions to learn the latent representation for the task of LP within the KG. This thesis focuses on proposing a model which takes the above described features into account and performs the task of LP i.e., head, tail prediction as well as triple classification. On the other hand, due to the structural differences amongst multiple KGs, their embedding spaces also exhibit different characteristics. Therefore, for the entity alignment task, these different vector spaces generated for different KGs are to be aligned to a single space to predict the missing links between the same entities across different KGs.

2 STATE OF THE ART

Link Prediction. So far, different KG embedding techniques have been proposed which can be categorized as translation based models, semantic matching models, models incorporating entity types, models incorporating relation paths, models using logical rules, models with temporal information, models using graph structures, and models incorporating information represented in literals and a detailed description of them are provided in [4], and [7]. Amonsgst them, he translational model [2] use scoring function based on distance and the translation is carried out with the help of a relation. GAKE [6] considers the contextual information by generating paths starting from an entity. On the other hand, DKRL [11] incorporates textual entity descriptions in the embedding model and uses TransE as the base model.

The textual entity descriptions present in the KGs provide information about the entity which might not be available otherwise in the KG. Also, the paths originating from an entity provide the structural contextual information about the neighboring entities. Therefore, in this thesis, paths and entity descriptions are modeled together to learn the embeddings of entities and relations for LP.

Entity Alignment. Entity Alignment is the task of aligning the same entities across different KGs. To do so, several embedding based methods have been proposed, in which a unified embedding space is learned using a set of already aligned entities and triples. A detailed description of these models for entity alignment is provided in [1]. The challenges of these models are: **(i)** They are supervised and require a set of aligned entities or triples as seeds for training. **(ii)** Some of the models require all the relations to be aligned between the KGs. However, in case of heterogeneous KGs which consist of different sets of relations, it is a challenging task to have a pre-aligned set of relations. **(iii)** The methods lack proper mechanisms to handle multi-valued relations. This thesis proposes an entity alignment model for heterogeneous KGs with multi-valued relations based on the unsupervised approach, i.e. without pre-aligned seeds for training.

3 RESEARCH QUESTIONS AND CONTRIBUTIONS

This section discusses the research questions and the corresponding contributions to address the challenges.

- *RQ1: Given an entity and a relation pair, how to predict the missing entity in a triple?*
  - The head or tail entity in a triple $< h, r, ? >$ or $<?, r, t >$ is predicted by defining a mapping function $\psi : E \times R \times E \rightarrow R$, where E and R are the set of entities and relations in the KG. A score is assigned to each triple where the higher the score of the triple indicates the more likely to be true.
- *RQ2: How to identify whether a given triple is valid or not?*
  - This is a triple classification task, in which a binary classifier is trained to identify whether a given triple is false (0) or true (1).

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Karlsruhe Institute of Technology, Institute AIFB, Germany email: russa.biswas@fiz-karlsruhe.de

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 69–70, Santiago de Compostela, Spain, August 29-30, 2020.

- *RQ3: How to predict the type information for an entity in a KG?*
  - Entity typing or Entity Classification is the process of assigning a type to an entity. To do so, different structural and literal information have been exploited to train a multi-label classification model for fine-grained entity typing.
- *RQ4: How to align the different embedding spaces of the KGs into a unified vector space to identify the* `owl:sameAs` *links?*
  - To align two different KG embedding spaces X and Y, a translation function τ coupled with a rotation function θ is introduced. The `owl:sameAs` links are then to be determined by vector similarity.

Therefore, the main contributions of this thesis are:

- A novel KG embedding model exploiting the structural as well as the textual entity descriptions in the KGs for head and tail prediction as well as triple classification.
- A neural network based multi-label hierarchical classification model for fine-grained entity typing using different features in the KG such as text and images along with the structural information.
- A novel translational model to align the different KG embedding spaces to identify the `owl:sameAs` links across multiple KGs.

4 LINK PREDICTION

To encapsulate the contextual information, random walks of 4 hops are generated starting from each entity in the KG. Predicate Frequency Inverse Triple Frequency (PF-ITF) [8] is used to identify the important relations for each entity. A sequence-to-sequence (seq2seq) learning based encoder-decoder model [10] is adapted to learn the representation of the path vectors in the KGs as shown in Figure 1. Given a path sequence, which is a combination of entities and the relations between them, such as $\{e_1, r_1, e_2, r_2, ..., e_n\}$, the input to the encoder is the corresponding embeddings (computed using TransE). These embeddings are passed through an attention based Bi-directional GRU which encapsulates the information for all input elements and compresses them into a context vector which is then passed through the decoder. A scaled dot product is employed as the attention mechanism. The representation of the textual entity descriptions is obtained using SBERT [9], followed by the same encoder-decoder model. ConvE [5] is used as a base model for scoring. For triple classification, the vectors are passed through a CNN model. Both triple classification and head/tail prediction are evaluated for FB15k and FB15k-237 datasets and the model outperforms the SOTA model DKRL as depicted in Table1. For the entity typing task, a multi-label CNN model is to be built.

5 FUTURE WORK

Entity Alignment. This task is yet to be addressed in this thesis. However, the basic idea is to adapt MUSE [3] which is an unsupervised multi-lingual word embedding alignment model to the KG alignment. A translation function coupled with a rotational function is to be used to align the related entities from different KGs. The same or related entities in different KGs will have overlapping information which could be exploited in an unsupervised manner.

ACKNOWLEDGEMENTS

This thesis is supervised by Prof. Dr. Harald Sack and Dr. Mehwish Alam.

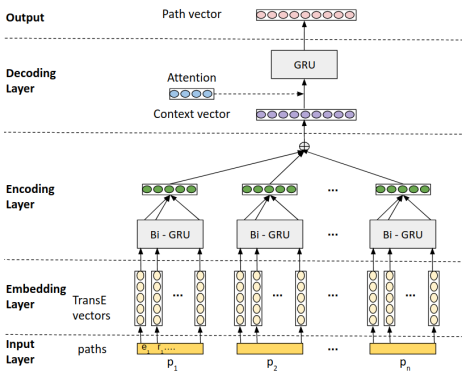

Figure 1. Encoder-Decoder Architecture

Table 1. Results on LP with FB15k and FB15k-237 datasets.

FB15k					
Models	MR	MRR	Hits@1	Hits@3	Hits@10
DKRL	85.5	0.311	0.192	0.359	0.548
Our model (w/o Attn.)	87	0.316	0.222	0.365	0.5615
Our model (w Attn.)	**85**	**0.335**	**0.243**	**0.383**	**0.59**
FB15k-237					
DKRL	90.5	0.298	0.187	0.337	0.523
Our model (w/o Attn.)	90.5	0.314	0.217	0.349	0.527
Our model (w Attn.)	**90**	**0.316**	**0.229**	**0.356**	**0.545**

REFERENCES

[1] Russa Biswas, Mehwish Alam, and Harald Sack, 'Is aligning embedding spaces a challenging task? an analysis of the existing methods', *arXiv preprint arXiv:2002.09247*, (2020).

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko, 'Translating Embeddings for Modeling Multi-Relational Data', in *NIPS*, (2013).

[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, 'Word translation without parallel data', *arXiv preprint arXiv:1710.04087*, (2017).

[4] Yuanfei Dai, Shiping Wang, Neal N Xiong, and Wenzhong Guo, 'A survey on knowledge graph embedding: Approaches, applications and benchmarks', *Electronics*, (2020).

[5] Tim Dettmers, Minervini Pasquale, Stenetorp Pontus, and Sebastian Riedel, 'Convolutional 2d knowledge graph embeddings (2018)', in *AAAI*.

[6] Jun Feng, Minlie Huang, Yang Yang, and Xiaoyan Zhu, 'GAKE: Graph aware knowledge embedding', in *COLING*, (2016).

[7] Genet Asefa Gesese, Russa Biswas, Mehwish Alam, and Harald Sack, 'A survey on knowledge graph embeddings with literals: Which model links better literal-ly?', *arXiv preprint arXiv:1910.12507*, (2019).

[8] Giuseppe Pirrò, 'Explaining and suggesting relatedness in knowledge graphs', in *ISWC 2015*.

[9] Nils Reimers and Iryna Gurevych, 'Sentence-bert: Sentence embeddings using siamese bert-networks', in *EMNLP-IJCNLP*.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, 'Sequence to sequence learning with neural networks', in *NIPS*, (2014).

[11] Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun, 'Representation Learning of Knowledge Graphs with Entity Descriptions', in *AAAI*, (2016).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 69–70, Santiago de Compostela, Spain, August 29-30, 2020.

Using Machine Learning for the Personalized Healthcare of Vascular Diseases

Ana Vieira[1]

Abstract. The growing number of people with vascular diseases has led to a burden of healthcare providers and an economic burden of patients. The development of personalized healthcare solutions has the potential to address these burdens while also promoting the patients' well-being. In this research work is proposed the use of machine learning algorithms to deliver personalized healthcare to patients with vascular diseases. By using machine learning to extract knowledge from patient data it is intended to improve clinical outcomes of patients while also supporting the health professionals in the clinical decision-making process.

1 INTRODUCTION

Vascular diseases affect a large percentage of the worldwide population, conditioning the patients' day-to-day lives. Due to their severity, these diseases require frequent follow-ups and regular hospitalizations. The follow-up consists of medical appointments where treatment plans are defined based on the patient's health condition. These treatment plans comprise the pharmacological intervention and activities that the patient needs to perform in order to control and improve their condition. However, this follow-up is only done sporadically, being difficult to monitor the patient's condition outside of healthcare settings. Due to the high prevalence of these diseases and the need of constant monitoring of the patient's condition, there is a growing burden on healthcare providers and an economic burden of patients. As such, it has been sought the research and development of technological solutions that tackle the current difficulties in the healthcare of vascular diseases.

The evolution of information technologies and their application in the health sector has opened the possibility for a change in healthcare, shifting from a traditional approached, based on the health professional's experience, to personalized healthcare, based on the patient's clinical information [5]. Personalized healthcare has the potential to reduce the current burdens of healthcare providers and patients while also promoting the patient's well-being.

There are already innovative solutions for the regular monitoring of the patient's health condition that generate high volumes of data, essential for the personalization of the healthcare of patients. However, this data is not processed [6], due to the lack of solutions focused on data analysis.

[1] GECAD, Institute of Engineering, Polytechnic of Porto, Porto, Portugal, email: aavir@isep.ipp.pt

Motivated by the current challenges that the healthcare of vascular diseases faces, this Doctoral work focuses on using patient data to deliver important insights about the patient's health condition. It is intended to improve patient outcomes and promote the patient's well-being by delivering personalized healthcare based on the analysis of patient data collected from monitoring sensors.

2 APPROACHES FOR PERSONALIZED HEALTHCARE

One of the main approaches that has been followed in the health sector for the personalization of healthcare is the use of recommendation systems. Being able of performing predictive analysis based on the patient's clinical data, these systems not only allow the support of the health professional but also of the patient. The literature presents recommendation systems for personalized healthcare aimed at both health professionals and patients [2, 8]. For health professionals, recommendation systems are used to support the clinical decision-making process, predicting the patient's condition and which treatment is the most appropriate. For patients, recommendation systems are used to monitor their physical activity, where the systems recommend appropriate exercises according to the patient's physical and clinical conditions. These systems are also being used to promote the patient's well-being and to present education information regarding the patient's health condition.

The analyses carried out by this type of systems use several intelligent methods, such as machine learning. Various machine learning techniques such as classification, clustering, and association algorithms, have been used in the literature to extract knowledge from patient data. Classification algorithms are mostly used for pattern detection, disease prediction, and detection of the patient's activity. Clustering algorithms are used to identify the disease and they are also used to predict it. Association algorithms are used in the clinical domain for the identification of relationships between diseases.

Although the application of information technologies has resulted in innovative solutions for the personalization of healthcare, these are still rarely used for the support of patients with vascular diseases [4]. Recent studies have sought to develop solutions for delivering personalized healthcare to patients with vascular diseases [1, 3, 7]. These solutions focus mainly on the monitorization of clinical information and on the presentation of personalized recommendations about physical exercises that the

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 71–72, Santiago de Compostela, Spain, August 29-30, 2020.

patient can perform, lacking functionalities regarding the disease management and the presentation of recommendations about which treatment to follow. In addition, the proposed solutions lack interfaces aimed to support the health professional in the clinical decision-making.

3 PROPOSED WORK

The main goal of the thesis is the research and development of algorithms capable of learning from patient data collected from monitoring sensors in order to support not only health professionals but also patients with vascular diseases. By doing so, it will be possible to deliver meaningful insights about the patient's health condition.

Specifically, it is intended to research and develop clustering and classification algorithms for the analysis of patient data. With the research of clustering algorithms, it is sought to identify the patient's health condition, based on the similarity with other patients. By grouping patients based on their similarity and analysing the evolution of their health condition it will also be possible to predict the patient's condition. By researching classification algorithms, it is intended to detect patterns and anomalies in the patient's condition, as well as predicting risks or benefits of medical interventions. Hybrid approaches to the analysis of patient data, i.e. the combination of classification and clustering algorithms, will also be applied to obtain better results. The insights generated by the learning algorithms will be made available to health professionals and patients through a recommendation system. This system will support health professionals in the clinical decision-making process by predicting possible reactions towards medical interventions, such as treatment or surgery, and recommending adequate treatments for patients. This system will also support patients through smart coaching. The recommendation system will promote the adoption of healthy behaviours and the patient's well-being by presenting personalized recommendations, adequate to the patient's condition, about activities that the patient can perform or behaviours that need to be adopted by the patient.

With the development of this thesis, it is intended to assess the impact of the insights generated by the machine learning algorithms on the clinical outcomes of patients and on the workload of health professionals.

4 ONGOING RESEARCH

The current research work being performed is the literature review of the topics that this thesis comprises:

1. Vascular diseases;
2. Main approaches for the analysis of health-related data obtained from monitoring sensors;
3. Approaches for personalized healthcare, including solutions in the domain of vascular diseases.

To identify and understand which disease-related parameters must be analysed, a detailed study of this thesis' clinical domain, vascular diseases, is being performed. This study is accompanied by a literature review of the main approaches used for the analysis of health-related data obtained from monitoring sensors, with particular focus on the identification of the patient's health condition and disease prediction. The study of the challenges that the analysis of health-related data faces is also included in this literature review.

The main approaches used for delivering personalized healthcare are also being studied, as well as an analysis of the current solutions available for vascular diseases. As a preliminary result of this review, mentioned above, it was found that information technologies are still rarely used for delivering personalized healthcare in the context of vascular diseases.

5 CONCLUSION AND FUTURE WORK

In this work a solution for the personalization of healthcare of patients with vascular diseases is presented, involving the use of machine learning algorithms, specifically clustering and classification algorithms, that will extract knowledge from patient data collected from monitoring sensors and will provide meaningful insights about the health condition to health professionals and patients. With the development of the Doctoral thesis, it is intended to improve the clinical outcomes of patients with vascular diseases while also reducing the workload of health professionals.

As future work of the thesis, the research and develop clustering and classification algorithms to process patient data will be carried out. The combination of these algorithms, hybrid approaches, will also be researched. The developed algorithms will be evaluated using adequate methods for the evaluation of machine learning algorithms and the best performing algorithms will be selected.

REFERENCES

[1] R. Ata *et al.*, "VascTrac: a study of peripheral artery disease via smartphones to improve remote disease monitoring and postoperative surveillance," vol. 65, no. 6, pp. 115S-116S, 2017.

[2] U. A. Bhatti, M. Huang, D. Wu, Y. Zhang, A. Mehmood, and H. Han, "Recommendation system using feature extraction and pattern recognition in clinical care systems," *Enterprise Information Systems*, vol. 13, no. 3, pp. 329-351, 2019/03/16 2019.

[3] N. Constant *et al.*, "A Smartwatch-Based Service Towards Home Exercise Therapy for Patients with Peripheral Arterial Disease," in *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2019, pp. 162-166: IEEE.

[4] M. E. Haveman *et al.*, "Telemedicine in patients with peripheral arterial disease: is it worth the effort?," (in eng), *Expert Rev Med Devices*, vol. 16, no. 9, pp. 777-786, Sep 2019.

[5] V. Jagadeeswari, V. Subramaniyaswamy, R. Logesh, and V. Vijayakumar, "A study on medical Internet of Things and Big Data in personalized healthcare system," *Health Information Science and Systems*, vol. 6, no. 1, p. 14, 2018/09/20 2018.

[6] A. K. Sahoo, S. Mallik, C. Pradhan, B. S. P. Mishra, R. K. Barik, and H. Das, "Intelligence-Based Health Recommendation System Using Big Data Analytics," in *Big Data Analytics for Intelligent Healthcare Management*, N. Dey, H. Das, B. Naik, and H. S. Behera, Eds.: Academic Press, 2019, pp. 227-246.

[7] A. Shalan, A. Abdulrahman, I. Habli, G. A. Tew, and A. Thompson, "YORwalK: Desiging a Smartphone Exercise Application for People with Intermittent Claudication," in *MIE*, 2018, pp. 311-315.

[8] R. Y. Toledo, A. A. Alzahrani, and L. Martínez, "A Food Recommender System Considering Nutritional Information and User Preferences," *IEEE Access*, vol. 7, pp. 96695-96711, 2019.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 71–72, Santiago de Compostela, Spain, August 29-30, 2020.

Adversarial Attacks on Android Malware Detection Models using Reinforcement Learning

Hemant Rathore[1]

Abstract. Malware analysis and detection is a rat race between malware designer and anti-malware community. Most of the current Android antivirus(s) are based on the signature, heuristic and behaviour based mechanisms which are unable to detect advanced polymorphic and metamorphic malware. Recently, researchers have developed state-of-the-art Android malware detection systems based on machine learning and deep learning. However, these models are prone to adversarial attacks which threaten the anti-malware ecosystem. Therefore in this work, we are investigating the robustness of Android malware detection models against adversarial attacks. We crafted adversarial attacks using reinforcement learning against detection models built using a variety of machine learning (classical, bagging, boosting) and deep learning algorithms. We are designing two adversarial attack strategies, namely single-policy and multi-policy attack for white-box and grey-box scenarios which are based on adversary's knowledge about the system. We designed the attack using Q-learning where a malicious application(s) is modified to generate variants which will force the detection models to misclassify them. The goal of the attack policy is to convert maximum Android applications (such that they are misclassified) with minimum modifications while maintaining the functional and behavioural integrity of applications. Preliminary results show an average fooling rate of around 40% across twelve distinct detection models based on different classification algorithms. We are also designing defence against these adversarial attack using model retraining and distillation.

1 INTRODUCTION

Today Android Operating System (OS) runs on more than 70% of the world's smartphone and holds the monopoly in the smartphone marketplace [8]. Android OS domination is due to its open-source nature, availability of a large number of apps, multiple third-party App Stores etc. [9]. The above factors have also made Android smartphones an obvious target of malware attackers. Malware is a software program designed with malicious intent against any target [10]. *Trojan-SMS.AndroidOS.FakePlayer.a* (2010) was the first malware detected on the Android OS. Since then, there has been exponential growth in malware detected on Android OS from $9,158$ in 2011 to 22.48 million in 2018 [3]. According to McAfee Mobile Threat Report, an average 8 million new Android malware was detected in each quarter of the year 2019 [7]. Antivirus industry (Norton, McAfee, Trend Micro, etc.) and anti-malware researchers provide the first line of defence against any malware attack [10]. Today most of the antivirus heavily rely upon signature, heuristic and behaviour based malware detection engines. These engines are exceedingly human dependent, time-consuming, non-scalable etc. and thus cannot detect advanced polymorphic and metamorphic malware [10].

Recently researchers have explored Machine Learning (ML) and Deep Learning (DL) to build Android malware detection systems which have shown promising results [10] [2]. Developing these systems is a two-step process: Feature Engineering and Classification. Feature engineering requires domain knowledge to extract various features using static/dynamic analysis. Arp et al. perform static analysis of $5,560$ Android apps to extract the features like permission, intent, API calls etc. and proposed a malware detection model using support vector machine which achieved 93.90% accuracy [1]. Li et al. developed SigPid using association mining to identify 22 most significant permissions for Android malware detection [6].

Machine learning and deep learning models are prone to adversarial attacks. Goodfellow et al. showed that a small number of intentional perturbations could force the classification model to misclassify with high confidence [4]. Earlier adversarial fooling was assumed to be due to non-linearity and overfitting of the models which was proven to be incorrect. Recently Yuan et al. proposed GAPGAN which perform a black-box attack on DL-based windows malware detection system using generative adversarial networks [11]. Kurakin et al. showed that adversary without any knowledge about the ML system can still fool the classification model [5]. The knowledge can consist of information about the dataset, feature info, model architecture and classification algorithm. In white-box attack, the adversary is assumed to have complete knowledge about the system, whereas in black-box attack the adversary does not know the system. Also, we can define the grey-box attack where an adversary has only limited knowledge which does not contain any information about model architecture and classification algorithm. Malware designers can also design and develop similar adversarial attacks on Android malware detection system, which will threaten the complete antivirus ecosystem. **Following are the specific contributions, the problem we are currently investigating and the direction of future work:**

- We propose the Single Policy Attack (SPA) for white-box scenario which uses Q-learning based on Reinforcement Learning (RL) to generate adversarial samples. Initial investigation of adversarial attack using SPA has shown an average fooling rate of more than 40% across twelve different Android malware detection models.
- We are also designing Multi-Policy Attack (MPA) for the grey-box scenario where an adversary has limited knowledge about model architecture and classification algorithm. The current design uses multiple Q-tables to optimize the adversarial attack with limited time and space complexity. Based on the current design, an average fooling rate of around 25% is achieved across twelve different Android malware detection models.

[1] Department of CS & IS, Goa Campus, Birla Institute of Technology and Science, Pilani, India, email: hemantr@goa.bits-pilani.ac.in

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 73–74, Santiago de Compostela, Spain, August 29-30, 2020.

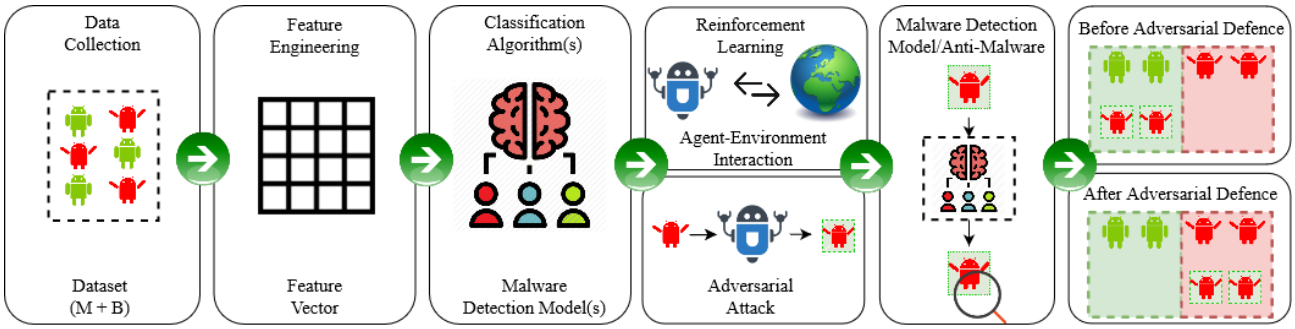

Figure 1. Overview of the proposed architecture for adversarial attack and defence

- The adversarial defence mechanisms should be able to defend against SPA and MPA, which are planned as future work based on model retraining and distillation. The final goal of this work is to propose an Android malware detection system using RL, which is more robust against adversarial attacks.

2 PROPOSED ARCHITECTURE

In this work, we will first act as an adversary to find vulnerabilities in the Android malware detection models. After a thorough investigation, we will design an adversarial defence to improve the robustness of the Android malware detection models against adversarial attack.

2.1 Problem Definition

Consider a dataset $D = \{(x_i, y_i)\} \in (X, Y)$ where X and Y represents an Android application and the corresponding class label (malware/benign) respectively. Feature vector x_i from X can be generated using static/dynamic analysis of application (malware/benign) and can be used to build Android malware detection models using different classification models. Performance metrics like accuracy, AUC-ROC, false-positive, false-negative etc. can be used to evaluate the quality of the detection models.

The goal of the adversarial attack on Android malware detection models is to modify maximum malicious applications (m) such that they are forced to be misclassified by the detection model(s). An adversary can assume white/grey/black box scenario to evade detection model(s) developed using a variety of classification algorithms (classical, bagging, boosting and DL). Another goal is to minimize the modifications in X or (x_i) and the changes should be syntactically possible, i.e. they should not disrupt the functional and behavioural aspect of the Android application.

2.2 Architecture Overview and Research Progress

Figure 1 illustrates the proposed architecture to develop robust Android malware detection models using adversarial attacks and defences. The process starts with data collection of Android applications (malware/benign). The next step is feature engineering which consists of feature extraction and feature selection. We perform static analysis of all the Android applications to extract features like permission, intent and API calls to be represented as a feature vector. In the third step, twelve different Android malware detection models were constructed using classical algorithms, bagging, boosting and DL algorithms. The fourth phase consists of two parts (1) In the first part RL agent interacts with the environment to generate an optimal policy (2) In the second part the optimal policy is used to modify the malware samples to be misclassified as benign. In the fifth phase, modified malware samples are tested with real-world Android malware detection models to calculate the fooling rate. Our preliminary results attain an average fooling rate of around 40% and 25% for permissions vector with SPA (white box scenario) and MPA (grey box scenario) respectively on twelve different detection models. We are still in the process of fine-tuning SPA and MPA. In the last phase, adversarial defence mechanisms are planned as future work and will be designed to improve the robustness of the detection models.

ACKNOWLEDGEMENTS

The author would like to thank Prof. Sanjay K. Sahay and team members (Piyush Nikam & Mohit Sewak) for their guidance and support.

REFERENCES

[1] Daniel Arp, Michael Spreitzenbarth, Malte Hubner, Hugo Gascon, and Konrad Rieck, 'Drebin: Effective and explainable detection of android malware in your pocket.', in *Network and Distributed System Security Symposium (NDSS)*, volume 14, pp. 23–26, (2014).

[2] Parvez Faruki, Ammar Bharmal, Vijay Laxmi, Vijay Ganmoor, Manoj Singh Gaur, Mauro Conti, and Muttukrishnan Rajarajan, 'Android security: a survey of issues, malware penetration, and defenses', *IEEE communications surveys & tutorials*, **17**(2), 998–1022, (2014).

[3] G DATA CyberDefense AG. Mobile Malware Report. `https://www.gdatasoftware.com`, (2019).

[4] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, 'Explaining and harnessing adversarial examples', *International Conference on Learning Representations (ICLR)*, (2014).

[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, 'Adversarial examples in the physical world', *International Conference on Learning Representations (ICLR)*, (2016).

[6] Jin Li, Lichao Sun, Qiben Yan, Zhiqiang Li, Witawas Srisa-An, and Heng Ye, 'Significant permission identification for machine-learning-based android malware detection', *IEEE Transactions on Industrial Informatics*, **14**(7), 3216–3225, (2018).

[7] Raj Samani. McAfee Mobile Threat Report. `https://www.mcafee.com/content/dam/consumer/en-us/docs/2020-Mobile-Threat-Report.pdf`, (Q1, 2020).

[8] Simon Kemp. Global Digital Report. `https://digitalreport.wearesocial.com`, (2018).

[9] Kimberly Tam, Ali Feizollah, Nor Badrul Anuar, Rosli Salleh, and L Cavallaro, 'The evolution of android malware and android analysis techniques', *ACM Computing Surveys (CSUR)*, **49**(4), 1–41, (2017).

[10] Yanfang Ye, Tao Li, Donald Adjeroh, and S Sitharama Iyengar, 'A survey on malware detection using data mining techniques', *ACM Computing Surveys (CSUR)*, **50**(3), 1–40, (2017).

[11] Junkun Yuan, Shaofang Zhou, Lanfen Lin, Feng Wang, and Jia Cui, 'Black-box adversarial attacks against deep learning based malware binaries detection with gan', *European Conference on Artificial Intelligence (ECAI)*, (2020).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 73–74, Santiago de Compostela, Spain, August 29-30, 2020.

Ranking Rules as a Tool for Reducing the Impact of the Distance in Machine Learning Methods

Noelia Rico [1]

Abstract. There are several machine learning methods that take decisions based on the distance between the objects in a dataset. How this distance must be measured is not always clear. Different distance measures may lead to different results and the information given by all of them is useful when the most suitable distance for the dataset is not defined. We propose to aggregate information given by different distances using ranking rules such as Borda Count.

1 INTRODUCTION

In the field of machine learning, there is a family of methods that performs their corresponding task based on the similarity between the objects in the data. This similarity is measured by means of a distance. By doing this, the method assumes that, given an object $\mathbf{x}$ and a dataset of objects $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, the objects in the dataset that are closer to $\mathbf{x}$ are more similar to $\mathbf{x}$ than the ones that are further. For both supervised and not supervised learning, methods based on this principles are quite popular, probably due to the fact that they are easy to use and also the results they generate are usually explainable [1].

How the distance between two objects should be computed is not very clear for all the scenarios as many definitions of distance exist [8]. Imagine that you are playing a game and you get to win the nearest object to your position. Which object is the nearest one? Intuitively, the distance is measured by humans using the Euclidean distance and thus, it seems natural that the object which is closest in a straight line would be chosen. On the other hand, someone could argue that, if the distance were measured, for example, with the Manhattan distance instead of the Euclidean distance, this object may or may not be the same, depending on how the other objects are situated.

Although for some problems the distance that better measures the similarity between the objects is clear, this is not the case for all the contexts in the real world. When performing a machine learning task, the dataset containing the input may combine data of different nature for which the best distance to measure the similarity between objects is not known *a priori*. In practice, the Euclidean distance is taken for granted as the distance that must be used, probably due to the fact that it is part of the standard distribution of the machine learning methods in most of the popular programming languages. Nevertheless, the results obtained by the method using this distance could be misleading if this distance happens to not be suitable for the input data [6].

This handicap emphasizes with the increasing popularity of machine learning in research areas not related with computer science, where methods are often used as a tool but without studying the principles that define how the method itself works. We aim to propose a standard implementation for some of the most common machine learning methods that aims to minimize the impact of the distance function in the results obtained. In our opinion, it could be interesting to propose a method that works with different distance functions so if the user does not know the suitable distance for the dataset at least the results can be more steady as they are obtained based on the consensus of the criteria given by multiple distances.

[1] University of Oviedo, Spain, email: noeliarico@uniovi.es

2 RANKING AGGREGATION

A new question issues from the fact that one distance is changed by multiple distance functions. How are the values given by the different distances merged? These cannot be simply aggregated with basic operations such as the average since not all the distance functions work with values of the same magnitude. In order to keep clear of the magnitude problem, objects from closest to furthest according to each distance can be rank to use the positions of the objects in the rankings to get a winner.

Suppose a contest where the four candidates c_1, c_2, c_3 and c_4 are judged by four different judges j_1, j_2 and j_3:

$$\begin{aligned} j_1&: \quad c_1 \succ c_4 \succ c_2 \succ c_3 \\ j_2&: \quad c_2 \succ c_4 \succ c_1 \succ c_3 \\ j_3&: \quad c_3 \succ c_4 \succ c_2 \succ c_1 \end{aligned}$$

In order to determine the winner of the contest, each of the judges gives a ranking of preferences from which the general winner must be elected. Given these rankings a question arises: which of the candidates should be the winner? At first sight, it is not clear whom should be elected, since each of the judges has the opinion that a different candidate is in the best position. Also, one of the candidates is in second position for all the judges, although it is never considered the best. Does it deserve to be the winner?

How to solve this problem has been studied for centuries now in the field of social choice theory. Scoring ranking rules propose a mechanism that given a set of rankings aims to select a winner ranking (from which a winner candidate can be deduced) assigning points to the candidates according to their position in the rankings. A popular scoring ranking rule of the many that have been proposed during the years is Borda count [2], which has been used in the field of social choice theory for centuries. When applying the Borda Count, each candidate is awarded one point every time that it is ranked at a better position than another candidate and half a point every time that it is tied at the same position as another candidate. All the points obtained by a candidate over all rankings are added up to obtain the

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 75–76, Santiago de Compostela, Spain, August 29-30, 2020.

score of the candidate. Finally, the Borda Count outputs the ranking of candidates from highest to lowest score. The scores obtained after applying this ranking rule to the contest, $c_1 = 3 + 0 + 1 = 4$, $c_2 = 1 + 3 + 1 = 5$, $c_3 = 0 + 0 + 3 = 3$ and $c_4 = 2 + 2 + 2$. Ranking the candidates according to their score the final value would be $c_4 \succ c_2 \succ c_3 \succ c_1$.

In our research we apply the presented concepts to extend machine learning methods based on distances by means of using different distances that give rankings of objects in the dataset. Then, those rankings are aggregated applying a ranking rule to the set of rankings and the winner ranking is the one employed in the different machine learning approaches we propose.

3 APPLICATION TO MACHINE LEARNING

The k Nearest Neighbours method (hereinafter kNN) [3] applied to classification problems is a classic method that aims to classify an object $\mathbf{x}$ into a *class* taking into account its most similar objects in the dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. This similarity is measured by the distance d between the objects, considering that the closest objects in the dataset are more similar than the ones which are further away. Thus, given the object $\mathbf{x}$ that needs to be classified, the other objects are ranked from closest to furthest and the k closest objects are selected, being k a value fixed prior to the beginning of the method. Then, the most common class among the k closest objects to $\mathbf{x}$ is established as the class of $\mathbf{x}$.

We propose to modify this method by introducing the criteria of multiple distances for determining the nearest neighbours. First of all, it is necessary to establish a set $\mathscr{D}$ that contains n different distances and also to fix the value of k. Once this has been done, in order to classify an object $\mathbf{x}$, using each of the distances $d \in \mathscr{D}$, the ranking of nearest objects to $\mathbf{x}$ that we aim to classify is computed. At this point, n different rankings are obtained. These n rankings are then aggregated using a ranking rule and the winner ranking is used to select the k objects in the best position and to classify the object $\mathbf{x}$ with the most common class among them. In [7] we study this idea using three different distance functions. Currently we are working in the study of how this method behaves in comparison with the classic $k$nn using a larger set of distances, more datasets and different ranking rules.

Similarly, we apply this idea to a clustering problem modifying the $k$means method. As proposed by MacQueen in the original paper [4] on $k$means, the goal of this method is to identify the partition of the dataset of objects $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ into k clusters that minimizes the distance between the points of the same group (known as total within sum of squares). To achieve this, first it is necessary to establish the number k of clusters the dataset is going to be divided into. Next, k initial centers are selected randomly. Once this has been done, the method of $k$means consists of two main parts, the *assignment* phase where the points are assigned to a cluster and the *update* phase where the center of the cluster is update. The algorithm iterates as follows: once the initial centers have been selected, the assignment phase occurs and the objects are assigned to the closest cluster (represented by the closest center). When all the objects have been assigned to a cluster, then the center of each cluster is recomputed as the centroid of all the points belonging to that cluster. These steps are repeated until all the points are assigned to the same cluster that in the previous iteration or a maximum number of iterations established beforehand is attained [5].

We propose a method where the assignation step is modified so, instead of using one single distance to determine the closest cluster, a set of distances $\mathscr{D}$ is used. Again, as we did for the modified version of $k$nn, we propose to combine the rankings of closest clusters obtained by each distance d using a ranking rule to assign the object $\mathbf{x}$ to a cluster.

1. Establish the number k of clusters the dataset is going to be divided into.
2. Select k initial random centers.
3. For each distance $d \in \mathscr{D}$, calculate the distance to all the centers and rank the centers from closest to furthest from $\mathbf{x}$ in a ranking r_d that is added to a set of rankings $\mathscr{R}$.
4. Using a ranking rule, aggregate all the rankings in the set $\mathscr{R}$ which contains all the rankings r_d given by the different distances.
5. Get the winner center from the first position of the resulting ranking after applying the ranking rule and assign the object $\mathbf{x}$ to the cluster it represents.
6. Recalculate the centers of the clusters.

These steps are repeated until none of the objects changes of cluster or a maximum number of iterations established beforehand is attained.

4 FUTURE WORK

Future work will cover the study of these methods applied in real world datasets, more concretely in computer vision problems. Furthermore, these techniques will be used to study the efficient evaluation of the clustering method proposed, adjusting clustering evaluation metrics that are also based on distances to use ranking rules. Once this has been established the ideas will be applied to other clustering methods that are not based in data partition.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli, 'Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda', in *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–18, (2018).

[2] Jean Charles Borda, *Mémoire sur les Élections au Scrutin*, Histoire de l'Académie Royale des Sciences, Paris, 1781.

[3] T. Cover and P. Hart, 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory*, **13**(1), 21–27, (1967).

[4] J. B. MacQueen, 'Some methods for classification and analysis of multivariate observations', in *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press, (1967).

[5] Laurence Morissette and Sylvain Chartier, 'The k-means clustering technique: General considerations and implementation in mathematica', *Tutorials in Quantitative Methods for Psychology*, **9**, 15–24, (2013).

[6] V. B. Surya Prasath, Haneen Arafat Abu Alfeilat, Ahmad B. A. Hassanat, Omar Lasassmeh, Ahmad S. Tarawneh, Mahmoud Bashir Alhasanat, and Hamzeh S. Eyal Salman, 'Distance and similarity measures effect on the performance of k-nearest neighbor classifier - a review', *Big Data*, (8 2017).

[7] N. Rico, R. Pérez-Fernández, and I. Díaz, 'Incorporating ranking rules into k nearest neighbours', in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, (2019).

[8] N Karthikeyani Visalakshi and J Suguna, 'K-means clustering using max-min distance measure', in *NAFIPS 2009-2009 Annual Meeting of the North American Fuzzy Information Processing Society*, pp. 1–6. IEEE, (2009).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 75–76, Santiago de Compostela, Spain, August 29-30, 2020.

Semantic Embeddings in Deep Convolutional Neural Networks

Armand Vilalta [1]

Abstract. One of the big questions in Artificial Intelligence (AI) is how can we represent knowledge. Here we focus on the field of transfer learning, specifically: how can we represent the knowledge learned by a Deep Neural Network for one specific problem, to be of use in different problems? We already found two important points: the contextualization of the knowledge learned and how a reduction of expressiveness can increase generalisation.

1 KNOWLEDGE REPRESENTATION

One of the big questions in Artificial Intelligence (AI) is how can we represent knowledge. The main challenge is that we need a representation we can use for "intelligent" computations. The question is not trivial at all and several answers have been provided over time according to which are the "intelligent" computations we wish to do over the data.

The paradigm of knowledge representation is ontology. It attempts to represent entities, ideas and events, with all their interdependent properties and relations, according to a system of categories. However, one of the main difficulties ontologies face in the field of AI is its creation. Ideally, we would like an AI to be able to learn new knowledge from data inputs, automatically avoiding the labour-intensive task of ontology creation. Important work has been devoted to this goal in the fields of ontology extraction, ontology generation or ontology acquisition. However, most of these methods obtain knowledge from natural language texts. They do not create knowledge from the data, but from a text describing the data. They do not "learn" the knowledge; they translate it. If we aim for a useful AI, capable of dealing automatically with the humongous sources of information available today, we definitely need automated ways to create and represent knowledge.

Since its popularization in 2012, when the Convolutional Neural Network (CNN) Supervision proposed by Alex Krizhevsky et al. [2] won by a fair margin the ImageNet Large Scale Visual Recognition Challenge [4], Deep Neural Networks (DNNs) have become the State-of-the-Art for Computer Vision (CV). They have been successfully applied to many other AI fields like text, voice, games or robotics and are applied to a myriad of contexts like web search, self-driving cars, health-care or finance. Deep Neural Networks have allowed AI to achieve a human performance level in many of the tasks and problems tackled so far. The power of DNNs relies on the internal representations of the information they learn from data. Each neuron layer in a DNN produces a new representation of the data based on the previous layer's representation. These representations are learned automatically so that they are useful to solve a particular task. That is why we say that DNNs are actually representation learning techniques. Unfortunately, to learn this representation space, deep networks require a lot of data and a lot of computational effort, which reduces the number of problems these models can be directly applied to.

Deep Neural Networks represent the information in the activation's of the neurons that compose them. These activations are usually represented as vectors or, more generally, as tensors. But understanding the actual meaning of these representations has proved to be an elusive task. Ontologies or knowledge graphs are derived from natural language, so they have an easy interpretation. On the contrary, DNNs' representations are learned just from data; therefore we do not even know the "language" in which they represent the data. Moreover, there is little understanding on why do they work or why do not depending on each particular case. We can neither be sure whether they will generalize properly to new unseen samples or not, what is called a "wild" generalization. For all this, Deep Neural Networks have been seen as a kind of Machine Learning (ML) black-box.

2 TRANSFER LEARNING

Even without having a clear understanding of whats going on inside a DNN, there had been successful attempts to reuse DNN's internal representations for a new problem, different from the original source task the DNN has been trained for. This is generally known as Transfer Learning (TL) [3]. Within Deep Learning, the field of Transfer Learning has three main applications: improving the performance of a network by initializing its training from a non-random state, enabling the training of deep networks for tasks of limited dataset size, and exploiting deep representations through alternative machine learning methods. The first two cases, where training a deep network remains the end purpose of the transfer learning process, are commonly known as transfer learning for fine-tuning, while the third case, where the end purpose of the transfer learning does not necessarily include training a deep net, is typically referred as transfer learning for feature extraction.

Transfer learning for feature extraction is based on processing a set of data instances through a pre-trained neural network by doing a feed-forward pass, extracting the neural activation values to generate a new data representation which can then be used by another learning mechanism. This is applicable to datasets of any size, as each data instance is processed independently. It has a relatively small computational cost since there is no deep net training. And finally, it requires no hyper-parameter optimization or tuning of any sort. Significantly, the applications of transfer learning for feature extraction are limited only by the capabilities of the methods that one can execute on top of the generated deep representations and addresses the needs of

[1] Barcelona Supercomputing Center, Spain, email: armand.vilalta@bsc.es

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 77–78, Santiago de Compostela, Spain, August 29-30, 2020.

many industrial applications where labelled data availability is often scarce.

3 CONTEXTUALIZATION AND DISCRETIZATION

Given the state of the art, we consider how we can improve transfer learning making it more general. To do that, we reason on the two paradigms of knowledge representation presented. If we make a comparison with actual human intelligence, it is like if we would like our neurons to talk easy words instead of talking via billions of spikes, which we can not understand. But how did we evolve a natural language from billions of spikes? The process is not known yet. But there is an important clear point: we can not say all what we think. Therefore, there is an important information bottleneck when we use natural language to express an idea. This is one o the key ideas of this research.

Another key idea is the generalization capability of human languages. If I say the word "chair" you immediately imagine an object which we both would recognize as a chair, but for sure we did not imagine the same chair. Even though both objects imagined will comply with similar functions, and most likely will have a similar shape. Therefore, when we use the word "chair" to communicate, we do it successfully most of the times, even if we are in a different context. So, this simplification of our mental representations allows us to reuse them in different contexts helping to increase its generalization capabilities.

Given that, we aim to obtain a more general representation of DNNs knowledge obtaining a "semantic" embedding. First of all we consider the context of the features. We are reusing a CNN trained for a different purpose so we must keep in mind that they may take a new meaning in the new context. You may skim over these lines, but not as you skim your milk. So, we want to use the DNN activations' meaning in the context of the new target task. We achieve it with a simple feature standardization in the context of the target task.

The key part is to create a bottleneck in order to reduce the representation space and improve generalization capabilities. We consider the simplest embedding that can retain meaningful information. To do so, we discretize each standardized feature value to represent either an atypically low value (-1), a typical value (0), or an atypically high value (1) in the context of the dataset. This specific representation is chosen because of its clear and simple semantic meaning. In this regard, it can be divided in two elementary propositions: first, whereas if the feature is significant; and second, if it is significant is it by absence or presence? In language parallelism it would be if a "word" is present, and second, if it is present is it used straight or negated?

4 RESULTS

We have already tested these hypotheses on classification problems obtaining very good results [1]. One of the most appealing properties of our solution is its robustness to the use of ill-suited pre-trained models in image classification. This is the most common scenario in real-world settings, when there is rarely a large dataset similar to the target task to pre-train with. When a dissimilar source task is used to pre-train the model, the difference with the baseline grows from 2.2 % accuracy on average to a remarkable 11.4 %. On top of that, the methodology developed is parameter-free, capable of providing results out-of-the-box. We also studied the application of the referred methods to multimodal embeddings. In all cases, proposed image embeddings obtained superior performance to that of a single layer instance normalized embedding [5, 6].

REFERENCES

[1] Dario Garcia-Gasulla, Armand Vilalta, Ferran Parés, Eduard Ayguadé, Jesus Labarta, Ulises Cortés, and Toyotaro Suzumura, 'An out-of-the-box full-network embedding for convolutional neural networks', in *2018 IEEE International Conference on Big Knowledge (ICBK)*, pp. 168–175. IEEE, (2018).

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, 'Imagenet classification with deep convolutional neural networks', in *Advances in neural information processing systems*, pp. 1097–1105, (2012).

[3] Sinno Jialin Pan and Qiang Yang, 'A survey on transfer learning', *IEEE Transactions on knowledge and data engineering*, **22**(10), 1345–1359, (2009).

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., 'Imagenet large scale visual recognition challenge', *International journal of computer vision*, **115**(3), 211–252, (2015).

[5] Armand Vilalta, Dario Garcia-Gasulla, Ferran Parés, Eduard Ayguadé, Jesus Labarta, Ulises Cortés, and Toyotaro Suzumura, 'Full-network embedding in a multimodal embedding pipeline', *arXiv preprint arXiv:1707.09872*, (2017).

[6] Armand Vilalta, Dario Garcia-Gasulla, Ferran Parés, Eduard Ayguadé, Jesus Labarta, E Ulises Moya-Sánchez, and Ulises Cortés, 'Studying the impact of the full-network embedding on multimodal pipelines', *Semantic Web*, **10**(5), 909–923, (2019).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 77–78, Santiago de Compostela, Spain, August 29-30, 2020.

Reinforcement Learning for Saturation-Based Theorem Provers

Bartosz Piotrowski[1]

Abstract. We describe a research project aiming at improving existing saturation-based automated theorem provers with reinforcement learning methods. These methods are meant to substitute heuristics governing certain aspects of the proving algorithms. We believe that there are hidden regularities appearing in the proving process which may be learned and exploited by the reinforcement learning agent to improve efficacy of the provers. The setup we consider aim at achieving balance between the speed of the prover and level of freedom for the reinforcement learning agent.

1 GOAL

The objective of our research project is to use reinforcement learning methods to improve state-of-the-art saturation-based automated theorem provers (like Vampire [3] or E [6]). The improvement should be demonstrated by increasing proving performance on multi-domain, broad, first-order theorem proving benchmarks like TPTP [7]. The designed methodology should not rely on consistent naming of symbols and formulae (which is the case for large-theory proving benchmarks like, e.g., MPTP [9] or HOList [1]).

2 MOTIVATION

Saturation-based theorem provers are currently the strongest available automated provers, as can be seen in the results of the CASC competition [8]. This kind of provers though, require a lot of *ad hoc* manual tweaking of their parameters. Moreover, there is a lot of incomprehensible "noise" in their runs. Machine learning methodology may be an effective remedy for dealing and optimizing complex behaviour of the saturation provers. However, it is unclear, which of the possible ways of applying machine learning to the provers would be the most effective. One of the issues which must be addressed is the trade-off between the speed of the run of the prover's algorithm and it's "intelligent behaviour". In our work, we propose an approach of applying reinforcement learning to state-of-the-art saturation-based provers addressing this trade-off.

Saturation-based automated theorem provers maintain two sets of clauses: the *processed* set and the *unprocessed* set. At the beginning, all the input axioms and the negated conjecture are clausified and put to the unprocessed set. Then, the saturation loop is run: from the unprocessed set one clause is selected and all possible inferences with the use of this clause and the processed clauses are performed. The selected clause is put to the processed set, and the newly inferred clauses are put to the unprocessed set.

[1] University of Warsaw and Czech Technical University

The selection of a clause to process is typically governed by manually-created heuristics. The Vampire prover, for instance, maintains two queues of unprocessed clauses. The first queues is ordered by "age" of the clauses, the second by "weight" (i.e., size) of the clauses. The clauses from the unprocessed set are selected for processing by selecting a queue and then the top clause from it. The queues are selected in weighted round robin fashion. The weights for the queues are fixed for the whole run. Their values, which influence the proving performance significantly, are pre-set manually on a basis of experience and conventional wisdom.

One may think of learning these weights with machine learning. Moreover, these weights do not need to be fixed. We hope it is possible to train a reinforcement learning agent, which dynamically switches between the queues on the basis of the current proof state. This may be even more interesting when a prover maintains more than two queues (which is the case, e.g, in the E prover).

Putting an RL agent in this place may be a good trade-off between the performance and the level of freedom for intelligent behaviour. Selecting each clause individually would be too slow. On the contrary, the queue may be selected for the next n iterations of the saturation loop and n may be suitably adjusted.

Why use reinforcement learning instead of supervised learning for the given purpose? Because the prover may arrive at the proof in (infinitely) many different ways, and there is no one best way. So no golden training data for supervised learning exists. Additionally, the environment induced by the prover is very noisy. So ultimately, reinforcement learning seems to be necessary if one wants to find a robust methodology to improve the state-of-the-art provers.

3 BASIC SETUP FOR REINFORCEMENT LEARNING

We propose the following basic setup for applying reinforcement learning for queue selection in the saturation provers.

The *state* available to the RL agent is a proof state in the prover characterized by a vector of manually selected statistics which may be computed efficiently. These statistics include: current sizes of the processed and unprocessed sets of clauses, and proportion of them; average length of a processed and unprocessed clause; the highest age in the unprocessed clauses; the lowest weight in the unprocessed clauses; elapsed time, remaining time. Other statistics of this kind are being considered.

The set of *actions* $\{0, \ldots, k\}$ available to the RL agent refers to selecting one of the available k queues for the next n iterations of the saturation loop, where n is a fixed hyper-parameter.

The *reward* given to the RL agent for its actions is 1 when a proof is found and 0 otherwise.

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 79–80, Santiago de Compostela, Spain, August 29-30, 2020.

The RL agent is trained by the classic policy-gradient algorithm – REINFORCE [10].

The training environment is initialized by selecting a training batch of TPTP problems. They are repetitively given to the prover as an input. For testing the trained RL agent, a separate testing batch of TPTP problems will be used.

The above-described setup was implemented for a simple saturation-based prover – PyRes[2]. The first experiments with this implementation are being run. This is meant as a preparation step before applying a similar setup in state-of-the-art provers like Vampire or E.

4 REFINING THE BASIC SETUP

One can think about multiple ways in which the described basic setup may be extended and refined. Below, we briefly show ideas we want to implement and evaluate experimentally.

The proof state may be characterized not by manually designed features/statistics, but with the use of graph neural networks, as presented in a recent work [4].

A non-zero reward may be given to the RL agent not only for finding a proof, but also for each queue-selection which resulted in picking up clauses which were actually used in the found proof. Additionally, the reward may depend on time and memory resources used by the prover.

Instead of using the basic REINFORCE policy-gradient algorithm, more advanced methods may be applied, including the actor-critic type of algorithms.

The training may be organized according to the curriculum learning idea – at the beginning, the RL agent may be trained on easier problems, and when achieving satisfactory performance, more difficult problems may be given.

5 RELATED WORK

There is few works in applying reinforcement learning to saturation-based provers.

In [5] authors present an idea where the RL agent learns to do several first inference steps, outside of the prover. This results in an augmented input for the prover; the reward is based on the success of the proof run and on the usage of resources by the prover.

In [2] authors present experiments in which the RL agent performs selections of the clauses in the saturation algorithm directly, without the use of the queues. The achieved performance in terms of the number of proofs found roughly matched the performance of the reference prover guided by standard heuristics. However, the paper does not state what was the time efficiency of the proposed solution. Likely, the RL-guided prover was much slower in comparison to the heuristic-guided prover. Our proposal of applying the RL-guidance on the level of queues is meant to be a trade-off between time-efficiency and level of freedom of the RL agent.

6 ACKNOWLEDGMENTS

The described research project is conducted in collaboration with Martin Suda from the Czech Technical University, who provides valuable expertise and good ideas.

[2] The implementation is available at `https://github.com/BartoszPiotrowski/PyRes/tree/reinforce`

REFERENCES

[1] Kshitij Bansal, Sarah M. Loos, Markus N. Rabe, Christian Szegedy, and Stewart Wilcox, 'Holist: An environment for machine learning of higher order logic theorem proving', in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, eds., Kamalika Chaudhuri and Ruslan Salakhutdinov, volume 97 of *Proceedings of Machine Learning Research*, pp. 454–463. PMLR, (2019).

[2] Maxwell Crouse, Spencer Whitehead, Ibrahim Abdelaziz, Bassem Makni, Cristina Cornelio, Pavan Kapanipathi, Edwin Pell, Kavitha Srinivas, Veronika Thost, Michael Witbrock, and Achille Fokoue, 'A deep reinforcement learning based approach to learning transferable proof guidance strategies', *CoRR*, **abs/1911.02065**, (2019).

[3] Laura Kovács and Andrei Voronkov, 'First-order theorem proving and Vampire', in *CAV*, eds., Natasha Sharygina and Helmut Veith, volume 8044 of *LNCS*, pp. 1–35. Springer, (2013).

[4] Miroslav Olsák, Cezary Kaliszyk, and Josef Urban, 'Property invariant embedding for automated reasoning', *CoRR*, **abs/1911.12073**, (2019).

[5] Michael Rawson and Giles Reger, 'Reinforcement-learned input for saturationprovers', *Proceedings of the 26thAutomated Reasoning Workshop*, (2019).

[6] Stephan Schulz, 'System description: E 1.8', in *LPAR*, eds., Kenneth L. McMillan, Aart Middeldorp, and Andrei Voronkov, volume 8312 of *LNCS*, pp. 735–743. Springer, (2013).

[7] G. Sutcliffe, 'The TPTP Problem Library and Associated Infrastructure. From CNF to TH0, TPTP v6.4.0', *Journal of Automated Reasoning*, **59**(4), 483–502, (2017).

[8] Geoff Sutcliffe, 'The CADE-27 automated theorem proving system competition - CASC-27', *AI Commun.*, **32**(5-6), 373–389, (2019).

[9] Josef Urban, 'MPTP 0.2: Design, implementation, and initial experiments', *J. Autom. Reasoning*, **37**(1-2), 21–43, (2006).

[10] Ronald J. Williams, 'Simple statistical gradient-following algorithms for connectionist reinforcement learning', *Mach. Learn.*, **8**, 229–256, (1992).

Proceedings of the 1st Doctoral Consortium at the European Conference on Artificial Intelligence (DC-ECAI 2020), pages 79–80, Santiago de Compostela, Spain, August 29-30, 2020.

Author Index