Twelfth International Workshop Modelling and Reasoning in Context

Jörg Cassens, Rebekah Wegener, Anders Kofod-Petersen

IJCAI 2021, Montreal, Quebec, Canada

MRC 2021 took place 19-20 August 2021 at IJCAI 2021, the 30th International Joint Conference on Artificial Intelligence, the world's premier AI Research venue. Like so many events for the last year and a a half, the conference was affected by the ongoing COVID-19 pandemic. IJCAI 2021 was a virtual conference "in Montreal-themed virtual reality".

The call for papers for the workshop invited original submissions that were not previously published or accepted for publication elsewhere. At least three members of the program committee did review each submission. A review form directed committee members to evaluate submissions for appropriateness, technical strength, originality, presentation, and to provide an overall score.

These preliminary proceedings present the seven papers that were accepted for presentation at the workshop. The final proceedings will be made available after the event.

MRC always aims to bring together researchers and practitioners from different communities, both industry and academia, to study, understand, and explore issues surrounding context and to share problems, techniques and solutions across a broad range of areas. By working together we can get a better understanding of context to be able to model and formalise it, to make it computable and to work towards a human-centric contextual AI.

The organisers would like to thank all the authors for submitting their papers and the members of the program committee for their valuable review contribution.

Workshop website
mrc.kriwi.de

Hildesheim, August 2021
Jörg Cassens, Rebekah Wegener, Anders Kofod-Petersen

Workshop Chairs

- Jörg Cassens – Department of Computer Science, University of Hildesheim, Germany
- Rebekah Wegener – Paris Lodron University Salzburg, Austria *and* Audaxi, Sydney, Australia
- Anders Kofod-Petersen – Alexandra Institute, Copenhagen, Denmark *and* NTNU, Trondheim, Norway

Program Committee

- David Aha – Naval Research Laboratory, USA
- Juan Carlos Augusto – Middlesex University, UK
- Henning Christiansen – Roskilde University, Denmark
- Adrian Clear – Newcastle University, UK
- Martin Christof Kindsmüller – Brandenburg University of Applied Sciences, Germany
- Ilir Kola – Delft University of Technology, The Netherlands
- David Leake – Indiana University Bloomington, USA
- Ana Gabriela Maguitman – Universidad Nacional del Sur, Argentina
- Tim Miller – University of Melbourne, Australia
- Grzegorz J. Nalepa – AGH University, Kraków, Poland
- Harko Verhagen – Stockholm University, Sweden
- M. Birna van Riemsdijk – University of Twente, The Netherlands

Contents

Intrinsic, Dialogic, and Impact Measures of Success for Explainable AI

Jörg Cassens[1] , **Rebekah Wegener**[2]
[1]University of Hildesheim, 31141 Hildesheim, Germany
[2]Paris Lodron University, 5020 Salzburg, Austria
cassens@cs.uni-hildesheim.de, rebekah.wegener@sbg.ac.at

Abstract

This paper presents a brief overview of requirements for development and evaluation of human centred explainable systems. We propose three perspectives on evaluation models for explainable AI that include intrinsic measures, dialogic measures and impact measures. The paper outlines these different perspectives and looks at how the separation might be used for explanation evaluation bench marking and integration into design and development. We propose several avenues for future work.

1 Explanations

Explanations are foundational to social interaction [Lombrozo, 2006], and numerous different approaches to achieving explainability have been proposed recently [Adadi and Berrada, 2018; Arrieta *et al.*, 2019; Doran *et al.*, 2017].

Criticisms of current research trends include that "accounts of explanation typically define explanation (the product) rather than explaining (the process)" [Edwards *et al.*, 2019]. Another criticism is that explanations are currently largely seen as a relatively uniform and definable concept, and even systems that take user goals with explanation into account treat it largely on the system side of development [Biran and Cotton, 2017]. Despite this, a human centred [Ehsan and Riedl, 2020] perspective on explanation in artificial intelligence is not new [Shortliffe, 1976; Swartout, 1983; Schank, 1986; Leake, 1992, 1995; Mao and Benbasat, 2000]. For example, Gregor and Benbasat [1999] point out that different user groups have different explanation needs.

We have earlier construed contextualised explanations based on user goals [Sørmo *et al.*, 2005]. This has been used to integrate explanatory needs in the system design process [Roth-Berghofer and Cassens, 2005; Cassens and Kofod-Petersen, 2007]. However, we have represented explanation as a static object rather than a dialogic process. This includes the ability of the technical system to make use of explanations as well, at least as part of the theoretical model, even if not in practical applications.

In our understanding, both human and non-human actors in heterogeneous socio-technical systems (or socio-cognitive, [Noriega *et al.*, 2015]) can be senders and receivers of explanations [Cassens and Wegener, 2019]. For example, a human should be able to "explain away" recommendations made by a diagnostic system in order to enhance the future performance. While we currently focus on the opposite situation, e.g. an artificial actor explaining its choice of recommendations to the human user, frameworks for designing explanation-aware systems should be able to account for different flows of explanations, at least in principle and by extension.

In order to distinguish this from views that see the machine as only the explainer, not the explainee, we make use of the established term explanation awareness [Roth-Berghofer *et al.*, 2007; Roth-Berghofer and Richter, 2008]. Our working definition is as follows:

- **Internal View**: Explanation as *part of the reasoning process* itself.
  - *Example:* a recommender system can use domain knowledge to explain the absence or variation of feature values, e.g. relations between countries
- **External View**: giving explanations of the found solution, its application, or the reasoning process *to the other actors*
  - *Example:* the user tells said recommender system why he chooses an apartment in Norway despite the system suggesting one in Sweden

Semiotics and philosophy as well as the human and social sciences provide a rich basis for applications in explainable AI [Miller, 2018]. There is sufficient empirical and theoretical evidence that explanations are generated, communicated, understood and used in ways that are:

- **Dialogic,** as suggested e.g. by Leake Leake [1995],
- **Contextualised,** as required by e.g. Fraassen van Fraassen [1980], comprised of
  - *Context Awareness* (knowing the situation the system is in) and
  - *Context Sensitivity* (acting according to such situation) Kofod-Petersen and Aamodt [2006]; Kofod-Petersen and Cassens [2011]
- **Multimodal,** as argued for by e.g. Halliday Halliday [1978] and being
- **Construed by user interest,** as noted by e.g. Achinstein Achinstein [1983].

Given these foundations, can a semiotic model of explanation as a form of multi-modal dialogic language behaviour in context be used to generate contextually appropriate explanations by computational systems? There is an extensive body of research focusing on generating and using explanations in AI. Currently, what is lacking is:

1. A theory of the **dialogic process** rather than a monologic product
2. A **cohesive theory of explanation** that is:
   - *contextually appropriate* (e.g. fitting people, topic, mode and place),
   - *semantically appropriate* (e.g. recognised as an explanation)
   - *lexicogrammatically optimal* (best possible multimodal realisation)
3. A framework for integrating explanatory capabilities in the whole **software development life-cycle**, from requirements elicitation over design and implementation through to its use
4. A framework for **evaluation measures**.

We will focus on the last aspect in the remainder of this paper. Research in particular when it comes to measuring the actual effectiveness and efficiency of explanations given to users still seems fragmented. We propose to measure explainability along three lines of inquiry. **Intrinsic measures** deal with the question of whether the system at hand can generate explanations at all. **Dialogic measures** look at whether the system's output is seen as an explanation by the users. Finally, **impact measures** ask whether the explanation generated is of any use. These questions should help to elicit and formalise requirements for explanations as well as find ways to evaluate solutions that are operationalised sufficiently to enable making claims of explainability that can be tested against and to further comparisons between systems and iterations of systems.

Explanations are needed during the whole life cycle of applications, from initial requirements elicitation over design and development processes to using the final system. Therefore, it makes sense to look at frameworks for measuring efficiency and effectiveness of explanations in the context of whole development and life cycle management processes. While quality measurements for explanation could eventually enable a final system score (for benchmarking purposes [Zhan *et al.*, 2019]), development is a cycle and it is contextual, and the goal is to be able to build "better" systems through "better" development processes, where explanatory success is part of success metrics. Given existing requirements for transparency, such perspective on evaluating explanations can also be part of a regulatory framework for ethical AI [Cath, 2018; Coeckelbergh, 2020; Erdélyi and Goldsmith, 2018].

2 Evaluations

Within HCI, a plethora of different instantiations of human centred development processes exist (e.g. [Beyer and Holtzblatt, 1997; Carroll, 2000; Cooper *et al.*, 2014; De Ruyter and Aarts, 2010; Holtzblatt and Beyer, 2016], to name a few). We should consider principles and methods for (designing and evaluating) explainability as additions to existing tool kits, agnostic to their use in established design processes whenever possible (limited by different ontological commitments).

Evaluation is central to Human-Computer Interaction, or rather: evaluations are central since they typically form a cycle and cover a system at various stages. While (formative and summative) evaluations are a cornerstone for human centred design, "it is far from being a solved problem" [MacDonald and Atwood, 2013]. We are generally in need for evaluation processes that are suited for emerging types of applications [Poppe *et al.*, 2007] and for sustainable and responsible systems development [Remy *et al.*, 2018].

But even if current (usability) evaluation methods [Dumas and Salzman, 2006] may ultimately fall short in the context of XAI, they can at least inform first iterations of evaluation standards. In particular when used in combination with theories and models from other areas, such as linguistics [Cassens and Wegener, 2008; Halliday, 1978; Wegener *et al.*, 2008], psychology [Kaptelinin, 1996], the cognitive sciences [Keil and Wilson, 2000], or philosophy [Achinstein, 1983; van Fraassen, 1980].

In this short paper, we cannot explore these contributions in detail, but we will briefly outline a tripartite model for capturing explanatory effectiveness that includes:

- **Intrinsic measures:** measures that pertain to the ability of a system to generate explanations.
  Can the system generate explanations?
- **Dialogic measures:** measures that pertain to interaction between the system and its users.
  Does the system's output work as an explanation for its users?
- **Impact measures:** measures that pertain to the potential, anticipated or actual impact of explanations.
  Is the explanation generated of any use?

We have separated these measures because each of these three types of measures has different methods for testing and they cover distinct aspects of what "explanatory success" can mean. It is only by combining these different perspectives that we can get a full picture of the explanatory performance of a system and the explanations that are a part of that system. While we can think of more perspectives, it is important to keep in mind that quality measures have to have a well defined scope and they need to be, indeed, measurable [Carvalho *et al.*, 2017]. Furthermore, for them to be able to improve processes in practice, they need to be sufficiently simple to apply.

2.1 Intrinsic Measures

These measure the ability of the system to generate explanations, both generally for the given context of use, but specifically the transparency and interpretability of the system itself or of aspects of the system such as ML models and data used as well as algorithmic and other design choices.

If a system or parts of a system are not transparent then it is unlikely to perform well on either dialogic or impact measures. We can think of intrinsic measures as a baseline for explainable AI – it is a necessary, but not sufficient condition. From a design process perspective, we will need to look at which components are necessary for explanation generation [Roth-Berghofer and Cassens, 2005]. Evaluating, we might explore the structure, modality and semantic characteristics of the different explanations to ensure that they are optimised for the situation. There are different specific methods that might be useful for intrinsic measures.

2.2 Dialogic Measures

Here we look at the question of whether that which has been generated actually works as an explanation to the user, in various conditions, situations and contexts. Under investigation is the shared semiotic process of explanation generator and explanation consumer. Different methods are going to be useful for dialogic measures including user studies, reaction studies, experimental studies and qualitative and quantitative methods in general. Explanations are inherently dialogic, so we are always going to want to know who is requesting the explanation, who is providing the explanation and how and why they are providing it. Tracking the exchange of information itself is a way to evaluate because it lets us see the reaction to the explanation.

Trustworthy AI could be an outcome of systems that score highly on dialogic measures. This does not mean that trustworthy systems will score well on impact measures, indeed, human and non-human agents are quite prepared to trust a system that may have negative impacts on their wellbeing. Trust can be engendered through a dialogically well performing malicious system and this is what makes impact measures so essential.

2.3 Impact Measures

Impact measures look at whether providing explanations offers benefits over the use of the system itself. These can be used both on an individual level and for larger systems.

For example, on the individual level, we might consider an adaptive learning system that offers explanations to further the learning goal [Sørmo *et al.*, 2005] a user might have. While dialogic measures can be used to evaluate whether such an explanation can function as an explanation to the student, it would remain unclear whether the explanation did actually improve learning outcomes.

These measures also look at the impact that the system can have in the world. How can it impact decisions, diagnoses, legal and access outcomes? The impact measures examine the potential, anticipated or actual impact of the system and the ability of the system to explain these repercussions to users in context. Here the concept of contextual AI is important because as Ehsan and Riedl argue, "if we ignore the socially situated nature of our technical systems, we will only get a partial and unsatisfying picture" [Ehsan and Riedl, 2020]. A good model of context is crucial for evaluating explanatory success [Kofod-Petersen and Cassens, 2007; Wegener *et al.*, 2008]. Ethical AI would be the outcome of a system that scores highly on impact measures. We would of course aim for beneficial and equitable AI, but ethical is at least a good baseline outcome. Here we might expect to see methods such as impact studies and hypothetical, scenario and risk modelling. It would be beneficial to know what the anticipated consequences of the explanation are for everyone involved.

3 Related Work

Mohseni *et al.* [2018] argue that the interdisciplinary nature of explainable artificial intelligence (XAI) "poses challenges for identifying appropriate design and evaluation methodology and consolidating knowledge across efforts". At the same time, this interdisciplinary approach is essential to the success of XAI. We view our suggestion as a way to complement, further consolidate, and operationalise their classification system for different goals in XAI.

Hoffman *et al.* [2018] propose a process model of explaining and suggest measures that are applicable in the different phases of their conceptual model. This compliments our (more abstract) notions of dialogic and (to a lesser degree) impact measures, whereas we see our notion of intrinsic measures as a prerequisite for their model. Both models can be systematically combined, depending on the need for granularity and aspects covered. Mueller *et al.* [2021] present some helpful higher-level psychological considerations that can serve as general templates for effective explanations.

Sokol and Flach [2020] introduce fact sheets with an extensive list of properties for different explanatory methods. This is complimentary to our approach and could be used to select methods supporting the measures chosen. A survey by Carvalho *et al.* [2019] on interpretability in machine learning is orthogonal to our model, with their results being useful for operationalisation of the intrinsic (e.g. their comparison of different methods) and the dialogic measures (e.g. the notion of explanation properties).

4 Conclusion

We propose a tripartite perspective on explanation in intelligent systems that aligns with (iterative and contextual) design and development processes of systems such that there is space for formative and summative evaluations. While it enables a final system score (which we propose for benchmarking purposes [Zhan *et al.*, 2019]), development is a cycle and it is contextual, and the goal is to be able to build "better" systems, where explanatory success is part of success metrics.

We have previously discussed the potential for Ambient Intelligence to be useful for creating explainable AI [Cassens and Wegener, 2019], particularly on the architecture level and with regard to capabilities subsumed [De Ruyter and Aarts, 2010]. We propose that the core characteristics and general architecture of ambient intelligent systems make them a good framework for developing XAI and that AmI systems themselves have the potential to become explanatory agents that can be mediators between humans and other systems. The concept of mediating explanatory instances has also been explored in the context of virtual explanatory agents [Weitz *et al.*, 2020] or as a user-specific "memory" of explanations [Chaput *et al.*, 2021].

Development of such mediators, concentrating explanatory capabilities in specialised agents that are contextually embedded in our surroundings and have the potential for personalisation and anticipatory interaction, could greatly benefit from a cohesive framework for measuring explanatory success from different perspectives.

References

Peter Achinstein. *The Nature of Explanation*. Oxford University Press, Oxford, 1983.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *arXiv preprint: 1910.10045*, 2019.

Hugh Beyer and Karen Holtzblatt. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, 2017.

John M Carroll. *Making use: scenario-based design of human-computer interactions*. MIT press, 2000.

Rainara Maia Carvalho, Rossana Maria de Castro Andrade, Káthia Marçal de Oliveira, Ismayle de Sousa Santos, and Carla Ilane Moreira Bezerra. Quality characteristics and measures for human–computer interaction evaluation in ubiquitous systems. *Software Quality Journal*, 25(3):743–795, 2017.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Jörg Cassens and Anders Kofod-Petersen. Explanations and case-based reasoning in ambient intelligent systems. In David C. Wilson and Deepak Khemani, editors, *ICCBR-07 Workshop Proceedings*, pages 167–176, Belfast, Northern Ireland, 2007.

Jörg Cassens and Rebekah Wegener. Making use of abstract concepts – systemic-functional linguistics and ambient intelligence. In Max Bramer, editor, *Artificial Intelligence in Theory and Practice II – IFIP 20th World Computer Congress, IFIP AI Stream*, volume 276 of *IFIP*, pages 205–214, Milano, Italy, 2008. Springer.

Jörg Cassens and Rebekah Wegener. Ambient explanations: Ambient intelligence and explainable ai. In Ioannis Chatzigiannakis, Boris De Ruyter, and Irene Mavrommati, editors, *Proceedings of AmI 2019 – European Conference on Ambient Intelligence*, volume LNCS, Rome, Italy, November 2019. Springer.

Corinne Cath. Governing artificial intelligence: ethical, legal and technical opportunities and challenges, 2018.

Rémy Chaput, Amélie Cordier, and Alain Mille. Explanation for humans, for machines, for human-machine interactions? In *WS Explainable Agency in Artificial Intelligence at AAAI 2021*, pages 145–152, 2021.

Mark Coeckelbergh. *AI ethics*. MIT Press, 2020.

Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. *About Face (fourth edition): the essentials of interaction design*. John Wiley & Sons, 2014.

Boris De Ruyter and Emile Aarts. Experience research: a methodology for developing human-centered interfaces. In *Handbook of ambient intelligence and smart environments*, pages 1039–1067. Springer, 2010.

Derek Doran, Sarah Schulz, and Tarek R. Besold. What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint: 1710.00794*, 2017.

Joseph S. Dumas and Marilyn C. Salzman. Usability assessment methods. *Reviews of Human Factors and Ergonomics*, 2(1):109–140, 2006.

Brian J Edwards, Joseph J Williams, Dedre Gentner, and Tania Lombrozo. Explanation recruits comparison in a category-learning task. *Cognition*, 185:21–38, 2019.

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

Olivia J. Erdélyi and Judy Goldsmith. Regulating artificial intelligence: Proposal for a global solution. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 95–101, New York, NY, USA, 2018. Association for Computing Machinery.

Shirley Gregor and Izak Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4):497–530, 1999.

Michael A.K. Halliday. *Language as a Social Semiotic: the social interpretation of language and meaning*. University Park Press, 1978.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint 1812.04608*, 2018.

Karen Holtzblatt and Hugh Beyer. *Contextual design: Design for life*. Morgan Kaufmann, 2016.

Viktor Kaptelinin. Activity theory: Implications for human-computer interaction. In Bonnie A. Nardi, editor, *Context and Consciousness*, pages 103–116. MIT Press, Cambridge, MA, 1996.

Frank C. Keil and Robert A. Wilson. Explaining explanation. In *Explanation and Cognition*, pages 1–18. Bradford Books, 2000.

Anders Kofod-Petersen and Agnar Aamodt. Contextualised ambient intelligence through case-based reasoning. In Thomas R. Roth-Berghofer, Mehmet H. Göker, and H. Altay Güvenir, editors, *Proceedings of the Eighth European Conference on Case-Based Reasoning (ECCBR 2006)*, volume 4106 of *LNCS*, pages 211–225, Berlin, September 2006. Springer.

Anders Kofod-Petersen and Jörg Cassens. Explanations and context in ambient intelligent systems. In Boicho Kokinov, Daniel C. Richardson, Thomas R. Roth-Berghofer, and Laure Vieu, editors, *Modeling and Using Context – CONTEXT 2007*, volume 4635 of *LNCS*, pages 303–316, Roskilde, Denmark, 2007. Springer.

Anders Kofod-Petersen and Jörg Cassens. Modelling with problem frames: Explanations and context in ambient intelligent systems. In Michael Beigl, Henning Christiansen, Thomas R. Roth Berghofer, Kenny R. Coventry, Anders Kofod-Petersen, and Hedda R. Schmidtke, editors, *Modeling and Using Context – Proceedings of CONTEXT 2011*, volume 6967 of *LNCS*, pages 145–158, Karsruhe, Germany, 2011. Springer.

David B. Leake. *Evaluating Explanations: A Content Theory*. Lawrence Erlbaum Associates, New York, 1992.

David B. Leake. Goal-based explanation evaluation. In *Goal-Driven Learning*, pages 251–285. MIT Press, Cambridge, 1995.

Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Craig M. MacDonald and Michael E. Atwood. Changing perspectives on evaluation in hci: Past, present, and future. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, page 1969–1978, New York, NY, USA, 2013. Association for Computing Machinery.

Ji-Ye Mao and Izak Benbasat. The use of explanations in knowledge-based systems: Cognitive perspectives and a process-tracing analysis. *Journal of Managment Information Systems*, 17(2):153–179, 2000.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv preprint: 1811.11839*, 2018.

Shane T. Mueller, Elizabeth S. Veinott, Robert R. Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J. Clancey. Principles of explanation in human-ai systems. In *WS Explainable Agency in Artificial Intelligence at AAAI 2021*, pages 153–162, 2021.

Pablo Noriega, Julian Padget, Harko Verhagen, and Mark D'Inverno. Towards a framework for socio-cognitive technical systems. In A. Ghose, N. Oren, P. Telang, and J. Thangarajah, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems X*, volume LNCS, pages 164–181. Springer, 2015.

Ronald Poppe, Rutger Rienks, and Betsy Dijk. Evaluating the future of hci: Challenges for the evaluation of emerging applications. volume LNCS 4451, pages 234–250, 01 2007.

Christian Remy, Oliver Bates, Jennifer Mankoff, and Adrian Friday. Evaluating hci research beyond usability. In *Extended Abstracts of the 2018 CHI Conference*, pages 1–4, 04 2018.

Thomas R. Roth-Berghofer and Jörg Cassens. Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In Héctor Muñoz-Avila and Francesco Ricci, editors, *Case Based Reasoning Research and Development – ICCBR 2005*, volume 3630 of *LNAI*, pages 451–464, Chicago, 2005. Springer.

Thomas Roth-Berghofer and Michael M Richter. On explanation. *Künstliche Intelligenz*, 22(2):5–7, 2008.

Thomas Roth-Berghofer, Stefan Schulz, David B Leake, and Daniel Bahls. Explanation-aware computing. *AI Magazine*, 28(4):122, 2007.

Roger C. Schank. *Explanation Patterns – Understanding Mechanically and Creatively*. Lawrence Erlbaum, New York, 1986.

Edward H Shortliffe. Computer-based medical consultations: Mycin. *New York*, 1976.

Kacper Sokol and Peter Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020.

William R. Swartout. What kind of expert should a system be? xplain: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21:285–325, 1983.

Frode Sørmo, Jörg Cassens, and Agnar Aamodt. Explanation in case-based reasoning – perspectives and goals. *Artificial Intelligence Review*, 24(2):109–143, October 2005.

Bas C. van Fraassen. *The Scientific Image*. Clarendon Press, Oxford, 1980.

Rebekah Wegener, Jörg Cassens, and David Butt. Start making sense: Systemic functional linguistics and ambient intelligence. *Revue d'Intelligence Artificielle*, 22(5):629–645, 2008.

Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. "let me explain!": exploring the potential of virtual agents in explainable ai interaction design. *Journal on Multimodal User Interfaces*, pages 1–12, 2020.

Jianfeng Zhan, Lei Wang, Wanling Gao, and Rui Ren. Benchcouncil's view on benchmarking ai and other emerging workloads. *arXiv preprint: 1912.00572*, 2019.

Explainable AI: a narrative review at the crossroad of Knowledge Discovery, Knowledge Representation and Representation Learning

Ikram Chraibi Kaadoud[1*], **Lina Fahed**[1], **Philippe Lenca**[1]
[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
{ikram.chraibi-kaadoud, lina.fahed, philippe.lenca}@imt-atlantique.fr

Abstract

EXplainable Artificial Intelligence (XAI) has recently become a very active domain, mainly due to the extensive development of black-box models such as neural networks. Recent XAI objectives have been defined in the state-of-the-art, for which specific approaches have been proposed. Implicit links can be found between XAI and other domains, especially related to knowledge and neural networks. We here aim to highlight these implicit links. We present a narrative review of research works in two domains: (i) Knowledge domain with focus on Knowledge Discovery and Representation, and (ii) Representation Learning. We discuss the similarity and joining points between these domains and XAI. We conclude that, in order to make black-boxes more transparent, XAI approaches should be more inspired and take advantage of past and recent works in Knowledge and Representation Learning domains. Through this paper, we offer an entry point to the domain of XAI for both multidisciplinary researchers and specialists in AI, as well for AI knowledgeable users.

Keywords: XAI, Knowledge Discovery, Knowledge representation, Representation learning, State representation learning, Manifold representation learning, Multi-view representation learning, Network representation learning

1 Introduction: XAI

During the last few years, eXplainable Artificial Intelligence (XAI), has become a very active domain[1] facing the high development of black-box models, such as neural networks [Guidotti *et al.*, 2018]. A new generation of XAI approaches have been proposed, for which several new concepts and terms are specific to application domains, data types or modeling. Application domains of XAI are multiple: machine learning, robotics, multi-agent systems, computer vision, Knowledge Representation and Reasoning, *etc.*

[Barredo Arrieta *et al.*, 2020] defined *"Given an audience, an explainable Artificial Intelligence is the one that produces details or reasons to make its functioning clear or easy to understand"*. Indeed, XAI aims to make Artificial Intelligence (AI) models more intelligible and accessible or to directly design explainable models and results [Buchanan and Shortliffe, 1984; Guidotti *et al.*, 2018; Barredo Arrieta *et al.*, 2020]. When the first case arises, XAI provides an *explanation* of the internal mechanisms and/or the reasons behind the AI model behavior *i.e.* its functioning and performance: an explanation is thus an interface between the AI model to explain and the target audience [Gunning, 2017]. We define an explanation as an information in a semantically complete format, which is self-sufficient and chosen according to the target audience regarding its knowledge, its expectations and the context. Hence, the purpose of an explanation is to clarify the cause, context and consequences of described facts through a set of statements or information [Van Fraassen, 1988].

It is important to underline that an explanation by its very nature is contextual: it is specific to a given target audience and also to a given context [Walton, 2004]. This makes XAI more challenging as automatic context understanding is still a very challenging task [Brézillon, 1999; Lim *et al.*, 2009; Augusto *et al.*, 2017; Hollister *et al.*, 2017] and no unified way for modelling context in intelligent environments has yet been proposed in the literature [Brenon *et al.*, 2018]. We emphasize that the context (*i.e* users context, goal context, *etc.*) is important to take into account in XAI. However, this point is not the focus of the paper.

In the state-of-the-art, an explanation can take different formats (*e.g.* visual, natural language, features relevance explanations, *etc.*) and combine several representations of the same information [Barredo Arrieta *et al.*, 2020]. Two main XAI techniques are proposed: (i) Ante-hoc techniques which consist in optimizing an already transparent AI model (*e.g.* linear regression, decision trees, *etc.*) by adding constraints or features in order to increase transparency through metrics, data visualisation, *etc.* (ii) Post-hoc techniques that aim to explain already built black-box AI models (mainly deep neural networks). Among famous Post-hoc techniques: LIME [Ribeiro *et al.*, 2016], SHAP [Lundberg and Lee,

*Contact Author

[1]We remind that Artificial Intelligence models with explanation goals have been questioned and investigated a long time ago such as in [Shortliffe, 1974]. However, the term XAI has been recently proposed.

2017], visual explanations, saliency mapping, *etc.* XAI has recently been covered by several reviews that reveal its complex and intrinsically multidisciplinary aspects from a technical, user or Human-Interaction viewpoint [Guidotti *et al.*, 2018; Gilpin *et al.*, 2018; Barredo Arrieta *et al.*, 2020; Vilone and Longo, 2020]. As examples, we can note technical-based reviews, as those related to reinforcement learning [Puiutta and Veith, 2020; Heuillet *et al.*, 2021], data-based reviews as those related to time series [Schlegel *et al.*, 2019; Rojat *et al.*, 2021] and application-based reviews related to healthcare [Adadi and Berrada, 2020] and banking [Burgt, 2020]. Other reviews are inspired by social science, human psychology, sociology or cognitive sciences [Miller, 2019; Capone and Bertolaso, 2020] in order to build ethical and fair models [Barredo Arrieta *et al.*, 2020].

One key issue that have not been discussed in the above cited reviews and that we would like to highlight, is the importance of knowledge in XAI. As an interface between an AI and a target audience, an explanation can be considered as an interpreter between the AI knowledge and the human target audience knowledge. Since knowledge domain is historical in AI, this raises in turn important questions about the impact of domains such as Knowledge Discovery and Representation on XAI. Furthermore, regarding black-box models and especially neural networks, it is important to mention that in recent papers, concepts like representation learning, knowledge/latent/hidden/abstract representation, latent space, *etc.* have been studied in order to tackle issues such as dimensionality, running time, algorithmic complexity, *etc.* However, to the best of our knowledge, no explicit relation has been defined between these concepts and XAI. We consider that as these concepts are increasingly recurrent in the literature, with no consensual definitions across fields, it becomes, in turn, more difficult to apprehend the XAI domain.

To address this issue, we propose a narrative review that, contrary to the above cited literature reviews, does not review XAI techniques. Our paper is a narrative review across several domains: a literature-based review that synthesizes technical research works related to domains that implicitly inspire XAI works. Our goal is to bring original insights, formulate new research questions and highlight promising future directions of XAI. More precisely, in this narrative review, we aim to address three questions. **First,** to centralize and clarify concepts recurrently used in AI domains but not always clear for XAI specialists. **Second**, to bring a new light to XAI by making explicit the links between XAI and two other domains: (i) Knowledge domain including Knowledge Discovery Process (KDP) and Knowledge Representation (KR), and (ii) Representation Learning (RL) more associated to deep learning domain. **Third**, to offer an entry point to the XAI domain for multidisciplinary or specialists in these domains.

Actually, these domains are often perceived as disconnected as most of the research is currently concentrated on only one of them [Sallinger *et al.*, 2020]. Despite this, we believe that it is important to enhance the links and the implicit relations that can be found between them. We thus consider that XAI has been indirectly inspired by these domains.

Figure 1 shows our vision as a schematic representation of XAI domain and both KDP, KR and RL domains. Table 1 lists the acronyms used. The paper is organized as follows: definitions are presented in section 2, KDP and KR in section 3, and RL in section 4. At the end of both last sections, we discuss the relation between the highlighted points, related to KDP, KR, RL and XAI. Finally, in section 5, we discuss future directions and perspectives related to XAI.

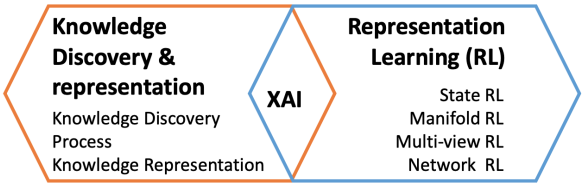


Figure 1: A schematic representation of XAI and its positioning at the crossroads of other domains.

Acronym	Research domain
XAI	eXplainable Artificial Intelligence
KDP	Knowledge Discovery process
KR	Knowledge Representation
RL	Representation Learning
SRL	State Representation Learning

Table 1: Acronyms of research domains discussed in this paper.

2 Definitions

This section is dedicated to the definition of several concepts related to Knowledge and Representation learning domains. Several definitions are inspired from state-of-the-art works.

Definition 2.1 The raw material that represents the input of an algorithm is called **data**. Data can be noisy, partial/complete, un/structured and of different types [Grazzini and Pantisano, 2015; Malhotra and Nair, 2015].

Definition 2.2 A data set is a collection of data that describes real-word **objects** (such as cars, documents, animal, *etc.*) through multiple properties called **features** [Bishop, 2006].

Definition 2.3 Once data is analyzed and correlated, it represents **information**. Information can be reproduced from data and its importance depends on the context it is generated from/for [Grazzini and Pantisano, 2015; Malhotra and Nair, 2015].

Definition 2.4 **Knowledge** is a set of information that is assessed by a human, *i.e.* human adds a value and semantics according to his/her own background and context [Grazzini and Pantisano, 2015; Malhotra and Nair, 2015].

Definition 2.5 In the data mining domain, a "***pattern*** *is an expression in some language describing a subset of the data or a model applicable to the subset*" [Fayyad *et al.*, 1996]. Hence, **Pattern extraction** designates the process of finding structures in data, fitting a model to data, or finding a high-level description of a data set.

Many data modeling approaches have been proposed in the state-of-the-art. We can cite reinforcement learning, graph-based approaches, neural networks, *etc.* We now define some important concepts related to these approaches.

Definition 2.6 Reinforcement learning is an approach in which an intelligent **agent** interacts with its environment through trial-and-errors actions in order to reach a goal. Each action leads to a modification of the **state** of the agent and the environment and the increase or decrease of a cumulative **reward** value. Actions are chosen according to a strategy that is called a **policy** [Barto and Sutton, 1995].

Definition 2.7 A **Manifold** is a topological structure of n-dimensions. For example, a one-dimensional manifold is a curve, a two-dimensional manifold is a surface, a three-dimensional manifold is a sphere.

Definition 2.8 A **network** is a collection of discrete objects called **nodes**, which are connected through **links**: it can be viewed as a graph with **vertices** and **edges**, both with attributes/weights or not [Fletcher *et al.*, 1991].

Definition 2.9 Neural networks are machine learning models with several architectures, that are usually structured by one or several layers (input, hidden and output). Each layer is composed of one or several computational **units** called artificial neurons - conceptually derived from biological neurons [McCulloch and Pitts, 1943; Abraham, 2005]. Computational units can also be a Long Short Term Memory (well known also as LSTM) [Hochreiter and Schmidhuber, 1997] or Gated recurrent units [Cho *et al.*, 2014]. A **deep neural network** have many hidden layers, units, and edges with weights. Units of layer n can be all or partially connected to units of layer $n+1$. Due to this inner complexity, deep neural networks are a typical example of **black-boxes**.

Definition 2.10 In neural networks, an **activation pattern** refers to units activation values of one of the layers. An activation pattern is a numerical **vector** of the size of the layer it is associated with. A **hidden pattern** refers to the activation pattern of a hidden layer.

In the literature of neural networks, concepts like latent space and latent representation have been developed and widely used. However, to the best of our knowledge, no complete definitions have been clearly proposed for such concepts. Due to the importance of both concepts in the rest of this paper, we choose to formulate their definition next.

Definition 2.11 Latent space refers to the abstract multi-dimensional space associated to each layer of a neural network where the representation of the learned data is implicitly built. Latent space contains the meaningful internal features (definition 2.2) representations of learned data, which makes it not directly interpretable. In a deep neural network (definition 2.9), each hidden layer, whether it has the same number of units or not, has its own latent space. It is thus possible to extract several implicit representations from this network. The latent space can be used to achieve a data dimensionality reduction, when the hidden layer is smaller than the input layer. This is the case for example with autoencoders and variational autoencoders [Kingma and Welling, 2014], models that can reduce high-dimensional inputs into efficient and representative low-dimensional representations [Roberts *et al.*, 2018b].

Definition 2.12 Latent or hidden representation refers to the data representation implicitly encoded by a neural network during the learning task and thus is hidden-layer-dependant [Bengio *et al.*, 2013]. It is a machine-readable data representation that contains features of the original data that have been learned by associated hidden layer. One key property of latent space (definition 2.11) is that real-world objects (definition 2.2) that are semantically close (*e.g.* cars of different brands), will end up grouped together in one latent space: their respective hidden representation in the corresponding layer, will be close to each other compared to other objects that are not semantically close (*e.g.* cats) [Roberts *et al.*, 2018a]. Thus, a latent representation is useful for pattern analysis (definition 2.5) and for similarity detection between objects (definition 2.2) using clustering methods.

3 Knowledge: discovery and representation

We now present two active research domains: KDP (section 3.1) and KR (section 3.2). Then, we discuss the relation between them and XAI in section 3.3.

3.1 Knowledge Discovery Process (KDP)

KDP is a human-centered domain that seeks useful knowledge (definition 2.4) through an iterative and interactive process that involves humans [Lenca, 2002; Cios *et al.*, 2007]. As the domains KDP, data mining, and Knowledge Discovery in Databases (referred to as KDD) are often used in a confused way, we consider that it is important to present a clarification about them, as follows:

- According to [Cios *et al.*, 2007], KDP and KDD designate the same process. However, KDP can be generalized to non-databases sources of data, while KDD emphasizes databases as a primary source of data.
- KDP and data mining are related to each other as well as to other domains like machine learning and statistics, but are clearly distinct. Indeed, according to [Fayyad *et al.*, 1996] and [Cios *et al.*, 2007], KDP is the global process of discovering useful knowledge from data, whereas data mining is a particular step within the KDP process that consists in applying algorithms to extract patterns (definition 2.5) or to build a model that fits the data.

There is no consensus about the steps of a KDP: nine steps in [Fayyad *et al.*, 1996], eight steps in [Anand and Büchner, 1998], six steps in [Wirth, 2000; Cios *et al.*, 2007] and five steps in [Cabena *et al.*, 1998]. However, we emphasize that globally KDP consists of three common main steps:

1. A pre-processing step for data collection or generation, data preparation, cleaning, curing, *etc.*
2. A data processing step where several techniques from statistics/machine learning/data mining, *etc.* communities can be used.
3. A post-processing step for visualisation, evaluation and validation.

At each step, the extracted information (definition 2.3) is usually evaluated by the human, given the context, to form knowledge[2] (definition 2.4). Thus, the target audience of the KDP is the human: application domain experts and decision makers. In addition, it is important to underline that two mains goals of KDP are usually defined [Fayyad *et al.*, 1996]: (i) verification of a user hypothesis, and (ii) discovery of valid and useful new knowledge that is understandable with respect to the data (definition 2.1) from which it is derived. These goals are thoroughly discussed in section 3.3.

3.2 Knowledge Representation (KR)

KR is a crucial question in AI [Malhotra and Nair, 2015]. Also known as "Knowledge Representation and Reasoning", KR aims at finding ways to efficiently structure specific domain knowledge for automated reasoning. In this way, intelligent machines can learn, draw inferences, make decision and answer questions related to this knowledge [Davis *et al.*, 1993; Shapiro, 2006; Davis, 2015]. Thus, seen in such a way, KR can be considered as a machine-oriented domain. The purpose of KR is neither about storing data, nor making actions but it is about allowing *"thinking by reasoning"* [Davis *et al.*, 1993]. Consequently, KR has been a key component for the conception of intelligent knowledge-based systems.

KR is also, according to [Malhotra and Nair, 2015], closely related to the Knowledge retrieval in the shape of ontologies (concepts for representing, storing and accessing knowledge [Guarino *et al.*, 2009]). KR techniques have also been widely developed and applied to semantic web [Hagedorn *et al.*, 2020], semantic networks [Malhotra and Nair, 2015], text interpretation and cognitive robotics [Davis, 2015]. In addition, from a user viewpoint, KR is important during the development of software systems in order to perform particular tasks, as well as for broader community of cognitive science whose goal is to constitute and organize knowledge from humans and machine perspectives [Das, 2003].

Knowledge Representation Learning (KRL)

KRL is the process of making AI algorithms model and learn a structured representation of domain-specific knowledge. As a consequence, concepts, relations between them and their representations can be encoded in a low-dimensional semantic space [Lin *et al.*, 2018]. For example, when knowledge is represented as a graph, the KRL process allows graph embedding and preserves semantic similarities [Xie *et al.*, 2018]. Notice that the development of deep learning algorithms and their performance on distributed representations (*i.e.* representations that describe features of the same data across layers) that reduce the computational complexity has contributed to the emergence of several KRL applications such as recommendation system [Zhang *et al.*, 2016], language modeling [Ahn *et al.*, 2016] and question answering [Yin *et al.*, 2016]. We consider that KR has recently become a more central domain in AI, and by extension in XAI. This is mainly due to the development of Representation Learning in neural networks (introduced in section 4).

[2]Notice that recent approaches like AutoML tend to perform all these steps automatically without user intervention [He *et al.*, 2021]

3.3 Discussion: relation between KDP, KR and XAI

We now discuss and highlight several links and common points between KDP, KR and XAI. As mentioned in sections 3.1 and 3.2, KDP is a human-centered domain, whereas KR is a machine-oriented domain. However, both domains are complementary: in KDP, the main question is "How to efficiently discover new or retrieve existing knowledge?", whereas in KR the tackled question is "How to represent the knowledge efficiently to be able to *reason* on it?".

It is important to highlight that both KDP and KR questions are also addressed and are crucial in XAI. Recall that the objective of XAI is to make the reasons behind AI behavior simple and accessible to a target audience regarding a given task and context. We consider that this XAI objective can be viewed and divided into two sub-objectives: (i) to discover the reasons behind AI behavior - which is the same as in a KDP problem -, (ii) to represent these reasons in a way that is intelligible for the human target audience, but also sometimes for an artificial one - which is the same as for a KR problem.

Let us first detail the links between KDP and XAI. In XAI, for black-boxes like deep neural networks [Guidotti *et al.*, 2018; Gilpin *et al.*, 2018], technical approaches are used to search the behavior of AI and make it explainable by providing an explanation that can take several forms and be multimodal [Barredo Arrieta *et al.*, 2020]. Explaining an AI model is therefore very inspired by KDP. The particular point is that in XAI, the input data (definition 2.1) is related to the black-box AI model. This input data can be of several types, *e.g.* activation patterns of hidden layers (definition 2.10), features or representations, and require the same techniques as in KDP.

Figure 2 represents a schematic representation of the transformation of data into knowledge, in KDP and XAI domains. It clarifies the similarities between both domains regarding the human intervention, and the role of the technical part, *i.e.* data mining and explainable methods.

Let us now go into deep details about knowledge representation in XAI. Two cases can be highlighted according to target audience: (i) human who uses the knowledge representation to reason and understand the situation, *e.g.* the decision maker and the application domain expert, depending on their expertise, role and goals, (ii) another AI system for which the input data is provided from a complex AI architecture.

Let us take an example in the domain of computer vision and especially classification using deep neural networks. Researchers have proposed approaches that exploit different AI algorithms and their latent representation (definition 2.12) as an input to the neural networks. The objective of such approaches is to perform both classification and explainability tasks through saliency masks applied to images and text generation [LeCun *et al.*, 2015]. This is one strategy among multiple others for the representation of knowledge in order to favor the explainability of the behavior of the initial AI model.

In addition, notice that KRL has been basically associated with deep learning algorithms, especially with techniques like graph representation learning [Hamilton, 2020] and concept learning [Dolgikh, 2018], which are both studied in the XAI domain [Xu *et al.*, 2018; Fazi, 2020].

Finally, it is important to highlight the importance of the target audience in both KDP/KR and XAI domains. Actually, the role of the target audience is decisive: knowledge is usually retrieved and shaped in order to answer a question of a target audience related to a given task and context such as verifying an hypothesis, inference and decision making. Knowledge representation and content in both KDP/KR and XAI domains are thus target and context dependant.

As a conclusion, XAI is closely related to both KDP and KR, and future works in XAI should take advantage of recent works in both domains, as well as older works.

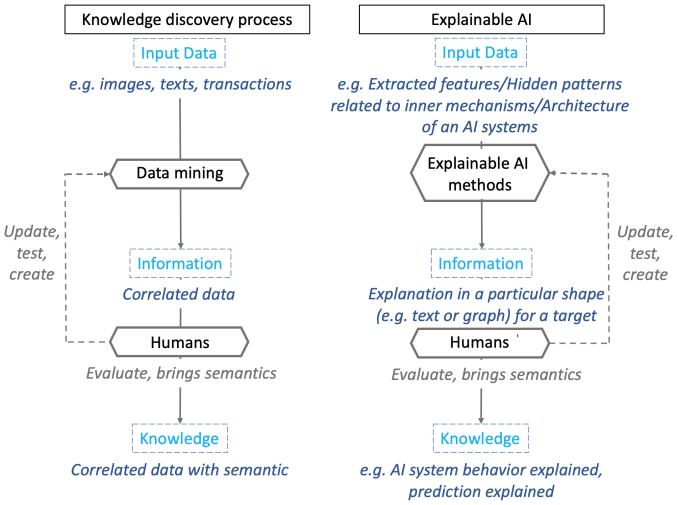


Figure 2: Two schematic ways for data transformation into knowledge: On the left, within a Knowledge Discovery process, and on the right within a XAI process.

4 Representation Learning (RL)

We now present the RL domain, its importance in deep neural networks, and RL sub-domains that are recurrent and popular.

4.1 RL introduction and definition

RL has been discussed as a key challenge related to different machine learning domains [Dietterich *et al.*, 2008] especially to neural networks. As first demonstrated by [Rumelhart *et al.*, 1986], in neural networks, back-propagation algorithms can generate useful internal representations of data in hidden layers. Since then, different approaches have been proposed in order to learn, analyze and visualize latent data representations [Gilpin *et al.*, 2018; Guidotti *et al.*, 2018]. Thus, RL has become an active research domain for which the objective is to study of latent representations in order to improve deep neural network efficiency [Bengio *et al.*, 2013].

RL - and synonyms like Data RL or Feature Learning [Zhong *et al.*, 2016] - focuses on *"learning representations of the data that make it easier to extract useful information when building classifiers or other predictors"* [Bengio *et al.*, 2013]. In other words, RL is designed to learn abstract features that characterize data [Lesort *et al.*, 2018].

RL algorithms can be classified into two categories: global and local RL algorithms. While the first ones tend to preserve the data global information in the learned feature space, the second ones focus more on preserving local similarity between data during learning the new representations [Zhong *et al.*, 2016]. Representations are not task-specific but are useful to machine learning algorithms to solve tasks, as well as to humans to comprehend the behavior of these last algorithms [Bengio *et al.*, 2013]. One of the reasons that makes RL popular is that representations express priors about the data. The expressed priors can vary within a single learning algorithm. Consequently, the characteristic of the priors variations leads to different RL approaches, that we classify into two categories: problems-oriented RL and data-oriented RL.

In the following section, we first present the concept of hierarchical representation in deep neural networks, a key property of RL. Then we present examples of particular cases of RL that are problems-oriented and data-oriented.

4.2 Hierarchical representations in deep neural networks

One key property of the RL domain in deep neural networks is the ability to provide both high level features and low level features for the same learned data. Recall that a deep neural network will encode a latent representation at each hidden layer (definitions 2.9, 2.12). Since the layer n units can be all or partially connected to the layer $n+1$ units, each layer uses the previous layer as input. If the previous layer is a hidden layer, then the input is already a latent representation, *i.e.* an abstract feature representation that characterizes the data. Thus, each layer extracts an abstract feature representation of the previous layer. As a result, a deep neural network learns multiple levels of abstraction and implicitly encodes a **hierarchy of latent and abstract representations** that are built progressively, layer by layer. The layers that are close to the input layer will encode a low-level feature representation, whereas those deeper inside the architecture will encode a high level feature representation. In other words, the closer the considered layer is to the output layer, the more the representation is abstract [Bengio *et al.*, 2013; Zhong *et al.*, 2016; Lesort *et al.*, 2018], as represented in Figure 3.

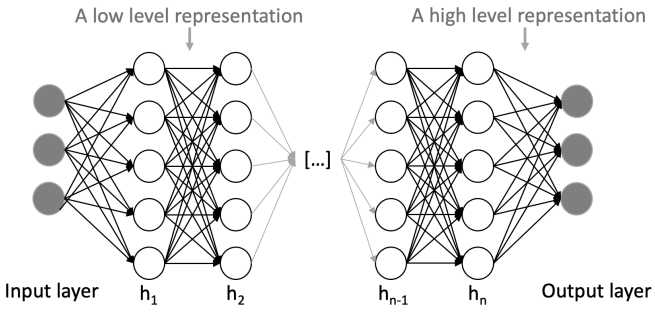


Figure 3: Illustrative and schematic representation of the position of a low level representation and a high level representation in a deep neural network. h_x refers to the x^{th} hidden layer in the network.

It has also been shown that, in deep learning algorithms, hidden representations tend to keep dominant information and propagate them across hidden layers, regardless the width

or depth increase of the deep neural networks [Nguyen *et al.*, 2021]. This characteristic of RL is also a key one for XAI: by extracting and comparing the low-level and the high-level representations of a deep architecture, we consider that it is possible to explicit the inner mechanism of the architecture by observing the differences between the representations. This will be discussed further in section 5.

4.3 Problems-oriented RL approaches

Recall that the objective of RL algorithms is to learn abstract features that characterize data. This objective can be challenging according the issues that one could face such as high dimensionnality of data or RL application to another AI paradigm like reinforcement learning (definition 2.6). In the following sub-sections, we describe two RL sub-domains: Manifold RL and State RL, that have recently shown great performances in deep learning and that deal with our core questions. The links with XAI are also briefly discussed.

Manifold RL

Manifold RL is particularly suited for dealing with high-dimensional data sets that are very difficult to visualize and less intuitive. However, within such data sets, data can locally belong to a subset that can be represented by a manifold. As stated in definition 2.7, a manifold is a topological structure of n-dimensions. Thus, Manifold RL corresponds to the learning of complex data representation in several dimensions while preserving the topological properties of the considered manifold. We consider Manifold RL as a non-linear dimensionality reduction approach, that can help to discover similarities in data for which dimensions have been reduced [Cayton, 2005; Bengio, 2009; Zhang *et al.*, 2011]. The Manifold RL domain aims at discovering manifold structure hidden in high dimensional data. It seeks to discover the intrinsic structure of a given manifold. Notice that when many manifolds are considered, we refer to this as multi-manifold RL [Lee *et al.*, 2016; Torki *et al.*, 2010]. It allows to both preserve the local geometric structure within distinct manifolds while ensuring the discriminability between them [Wu *et al.*, 2020].

When more neural networks transparency is required, the visualisation of latent representations is essential: it allows to develop an intuition about the distance between subsets of data represented by their associated latent manifold representations. Consequently, we consider that this dimension-reduction characteristic is therefore of great practical interest for XAI. Indeed, reducing the complexity due to the high dimensions can strongly contribute in understanding the inner mechanisms of models exploiting the data, but also the role of the data subsets on the models behaviors.

State RL (SRL)

In addition, RL can also concern domains where data are in a low dimensional space. SRL is *"is a particular type of representation learning that aims at building a low-dimensional and meaningful representation of a state space, by processing high-dimensional raw observation data (e.g., learn a position (x, y) from raw image pixels)."* [Heuillet *et al.*, 2021]. This domain is thus particularly suited for learning features in reinforcement learning, robotics and control scenarios. Thus, learning in SRL for an artificial agent is rather related to building a latent model of the environment and the task to perform through interactions [Lesort *et al.*, 2018]. In addition, it has been shown that SRL provides three main advantages for several research domains [Heuillet *et al.*, 2021]:

- The learned features are of low dimensions which improves speed and generalization of deep learning models [Lesort *et al.*, 2017].
- SRL helps improving performance in some reinforcement learning steps such as policy learning [Heuillet *et al.*, 2021].
- Learning representations of states (definition 2.6), actions or policies provide meaning to explain a reinforcement learning algorithms. Indeed, SRL allows to learn representations that capture the variation in the environment generated by the action of the agent [Lesort *et al.*, 2017; Heuillet *et al.*, 2021].

It has been shown that SRL is particularly suitable to make the behavior of an artificial agent and the reasons of this behavior accessible for humans [Lesort *et al.*, 2017; Lesort *et al.*, 2018; Heuillet *et al.*, 2021]. Consequently, we can consider SRL as an example of domains used for explanation goals in reinforcement learning.

4.4 Data-oriented RL approaches

In RL, several approaches tackle the problem of increasing data volumes, their heterogeneity and the multiplicity of their sources. We can consider them as data-oriented approaches and present two of them: the Multi-view RL and the Network RL. We also highlight the link between RL applied to real-world data-oriented problems and XAI domain.

Multi-view RL

In real-world applications, each object can be described by multiple features (definition 2.2) [Xu *et al.*, 2013]. It is thus referred as Multi-view data. These features, also referred to as views, constitute complementary and diverse information of the same data [Xu *et al.*, 2018]. For example, one information can be obtained through multiple sources, which is the case in the application where different people are talking about the same thing. Another example can be an image that is described via a set of visual features such as color, shape and textures. Multi-view RL is thus concerned with the problem of the integration of information from multiple views and uncovers the latent structure shared by multiple views, while preserving the original information and the global meaning [Zhu *et al.*, 2014; Xu *et al.*, 2018]. It has been shown that Multi-view RL can facilitate extracting useful information when developing prediction models [Li *et al.*, 2018] and also helps encoding concepts and semantics in deep neural network [Xu *et al.*, 2018]. Recently, Multi-view RL has been used to design an explainable recommendation system [Gao *et al.*, 2019], where authors claim that *"it is difficult to model the relationships between high-level and low-level features since they have overlapping meaning"*. To overcome this issue, a Multi-view learning approach has been proposed by considering different

levels of features as different views. The learned representation can then be a representation of different levels of features of the input data. Accordingly, we consider that Multi-view RL can be employed for explainability tasks.

Network RL

Network RL is a learning paradigm proposed to analyze networks such as graphs, and thus allows users to deeply understand the hidden features of graphs [Sun *et al.*, 2020]. This domain aims at learning in a low-dimensional space of network vertices (definition 2.8), while preserving the structure of the network topology, the content of the vertices and other information as vertices attributes and links attributes. Network RL can be considered as a dimensionality reduction technique and an intermediate step to solve a target task [Zhang *et al.*, 2020]. Since the information of the original network is preserved in a new vector-based representation, conventional vector-based machine learning algorithms can be applied. Thus, Network analysis and mining tasks become easier as there is no more need to use complex algorithms directly designed for graphs.

Consequently, Network RL has multiple applications such as: vertex classification, link prediction, clustering, visualization and recommendations [Dong *et al.*, 2020; Zhang *et al.*, 2020]. Network RL approaches have been widely applied to information networks [Sun *et al.*, 2020; Zhang *et al.*, 2020] and are becoming increasingly popular for capturing complex relationships in various real-world applications [Yang *et al.*, 2015; Sun *et al.*, 2020; Zhang *et al.*, 2020], such as social networks, citation networks, telecommunication networks, biological networks, recommender systems, *etc.*

In addition, Network RL is essential in the study of heterogeneous information networks (*i.e.* where vertices are of different types), in order to capture semantic proximity between vertices representations [Dong *et al.*, 2020]. Given the high scale of some networks that can range from hundred to billions of vertices and the heterogeneity of information, we believe that Network RL and XAI should be considered together in order to perform efficient and explainable analytical tasks. Also, in related applications, an in depth analysis using XAI techniques and Network RL can help interpreting empirical results and providing a deep understanding of the applied black-box model. To conclude, Network RL should be considered as a dimensionality reduction technique whenever graph-data structure is involved in the design of XAI.

4.5 Discussion: relation between RL and XAI

We have presented several research works in RL (Manifold RL, State RL, Multi-view RL and Network RL) and we next highlight common points between RL and XAI.

First, let us discuss the contribution of the **hierarchical RL** on XAI modeling. Recall that while RL focuses on learning a data representation in order to get a better performance of the AI model [Bengio *et al.*, 2013], XAI is interested in exploring this representation to explain the performance and behavior of the model. This representation varies according to the techniques used in the involved AI models (*e.g.* an artificial agent or a neural network). In the case of deep neural networks models, the hierarchical level of representations is important for XAI, as it allows to extract different types of information that can be used in several ways:

- The study of low-level representations can help to detect important features used by the deep network to make a prediction. This contributes to the explanation and understanding of the deep network by determining features involved in a particular output (*i.e.* a prediction).
- The study of high-level representations can help to detect groups of features involved in a prediction, and how and where a deep neural architecture deals with these groups. This is interesting to explain relevant hidden information and their location within the architecture.

For example, a hierarchical multi-scale deep recurrent network approach has been proposed for data sequences [Chung *et al.*, 2016]: in order to discover temporal dependencies in data, the latent hierarchical structure in the sequences has been exploited without using explicit boundary information. Accordingly, we consider that the hierarchical structure of the latent representations is an important characteristic of deep networks in order to propose a model-specific XAI modeling.

Second, we focus now on the contribution of **problems-oriented** and **data-oriented** RL approaches discussed above on the explainability of AI models.

- Recall that for high-dimensional data sets, **Manifold RL** allows to perform dimension reduction in the latent space while preserving the distance or similarities between data. Consequently, one of the main advantages is that visualisation of the data representation inside the latent space allows to get a better intuition and understanding of the inner mechanisms of models.
- Recall that in reinforcement learning, **SRL** allows to explicit the agent state changes while performing a task in a given environment. This is similar to the XAI objective as it makes the behavior of an artificial agent explicit and more intelligible for a given target audience. Also, recent works have mentioned that State RL can be viewed as a mean for XAI in reinforcement learning [Heuillet *et al.*, 2021]. Other works describe State RL as an approach for robotics and control scenarios that provides easier interpretation of the variation in the environment [Lesort *et al.*, 2017]. Consequently, we can consider that the goals of SRL are in line with those of XAI.
- Through the presentation of **Multi-view RL** and **Network RL** in section 4.4, we have shown that real-world applications of RL techniques that can be more specific to a particular data type or data organisation, are also linked to XAI. Indeed, an AI model can learn from multiple data sets of complex data representation such as networks (*e.g.* social network modeling, biological networks). The complexity of the learned data can also impact the behavior of the AI model. Consequently, this allows us to conclude that adopting RL approaches that take into account the type of learned data, is a way to make AI models more explicit and explainable.

Figure 4 summarizes the above conclusions and questions tackled throughout the section 4. Table 2 summarizes RL domains and some examples of application domains.

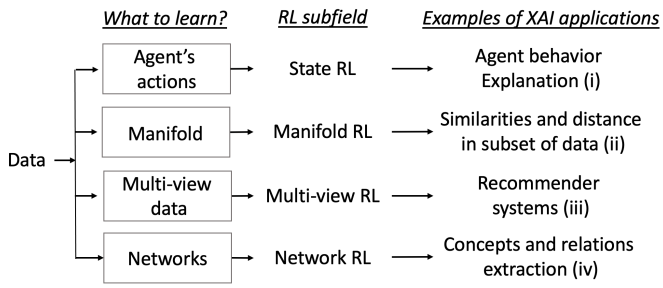


Figure 4: Questions addressed throughout the paper in section 4. Associated reference to each example: (i) [Madumal *et al.*, 2020], (ii) [Torki *et al.*, 2010], (iii) [Gao *et al.*, 2019], (iv) [Qi *et al.*, 2020]

Approach	Non-exhaustive examples of application domain
RL	Speech recognition [Liu *et al.*, 2020] Object recognition [Wang *et al.*, 2020] NLP [Mikolov *et al.*, 2013; Bérard *et al.*, 2016]
State RL	Robotics [Lesort *et al.*, 2017] Numerical artificial agent [Madumal *et al.*, 2020]
Manifold RL	Data mining [Torki *et al.*, 2010]
Multi-view RL	Concept learning [Xu *et al.*, 2018] Image processing [Su *et al.*, 2011] Recommender systems explainability [Gao *et al.*, 2019]
Network RL	Networks of concepts [Yang *et al.*, 2015; Qi *et al.*, 2020] Identification of genes in biology [Ietswaart *et al.*, 2021] Community detection in social networks [Tu *et al.*, 2018]

Table 2: A Summary of RL approaches, examples of application domains (NLP stands for Natural language processing).

5 Discussion and conclusion

We now summarize the highlighted points presented in previous sections. We also present promising directions related to the XAI domain. Since our paper is a multidisciplinary one at the crossroad of several domains, we have **first** (in section 2) centralized and clarified definitions of several concepts, that could indeed seem basic and well-known to involved AI experts, but are important to bridge the discussed domains. A special focus has been made on latent space, latent representation and hierarchical representation which are essential for knowledge extraction in deep neural networks and thus in XAI. To the best of our knowledge, no previous work has established a clear definition of these concepts for XAI community. This is necessary to allow the collaboration between the different domains necessary to build XAI. **Second**, we analysed and highlighted the existence of relations between Knowledge domains (KDP, KR), RL and XAI.

As we have shown in section 1, the goal of XAI is to convey the most semantically complete explanation to a target audience in order to answer a particular question within a given context. This explanation should take into account two important points: (i) the prior knowledge of the target audience regarding the application context, and (ii) the technical aspects of the AI used model that provided solutions to a specific task, and that thus contributed, due to its complexity/opacity, to the emergence of the question behind the need of XAI, *i.e* in short, "What are the reasons behind the results and/or how the AI model reaches these results?".

We consider that XAI is technically at the crossroad of at least two domains: (i) KDP and KR when viewed from a human perspective, and (ii) RL that tackles implicitly the same objectives as XAI, from a technical and algorithmic perspectives. KDP, KR, RL domains, while distinct, are overlapped. They do and should have an explicit impact on XAI approaches:

- First, as we have previously mentioned, several XAI approaches are indirectly inspired by the domain of Knowledge (KDP, KR and data mining) as both tend to express information from data. However, it is important to recall that, in XAI the input data reflects the internal mechanisms of the AI model, its predictions, and/or its behavior. The evolution of the Knowledge domain is therefore an inspiration area for XAI.
- Second, the development of AI approaches and in particular of deep learning, has blurred the boundaries between KR and RL, since several KR approaches involve RL and deep learning. In addition, recall that while RL is interested in features modeling for algorithmic issues (performance, dimensionality, *etc.*), XAI is interested in features since it contributes to explicit the inner mechanisms behind the results. This implies that KR, RL and XAI are indeed interested in the data representation in order to answer different but related questions. We thus consider that, in order to make a significant progress, XAI future works should not forget KR and RL past and recent works as inspirations.

KDP, KR and RL have been extensively confronted with, first, issues related to providing a data-driven explanation to different stakeholders according to their expectations and context, and second, issues related to biases and fairness in AI [Nelson, 2019]. This highlights the human significant role on data processing and bias detection in AI towards XAI. We believe that this review is all the more topical and important as works about the alliance between symbolic AI and connectionist AI should be more and more important in the next years[3], *e.g.* injecting a priori knowledge into neural networks to limit unethical AI [Goebel *et al.*, 2018] and biases [Gordon and Desjardins, 1995; Leavy, 2018; Lepri *et al.*, 2018; Nelson, 2019]. We are convinced that very promising directions can be taken in XAI future works by taking advantage of KDP, KR and RL development to help design ethical, unbiased and human-centered XAI. To conclude, we point out that other domains, not discussed in this paper, also impact XAI directions such as cognitive psychology [Le Saux *et al.*, 2002], cognitive sciences for biases studies [Soleimani *et al.*, 2021], social sciences [Miller, 2019] and Human Machine Interaction field [Le Saux *et al.*, 1999; Mueller *et al.*, 2021; Ehsan *et al.*, 2021].

6 Acknowledgments

Thanks to the *Conseil régional de Bretagne* and the *European Union* for funding this work via the *FEDER* program.

[3]The alliance of symbolic AI and connectionist approaches have been proposed a long time ago, *e.g.* [Honavar, 1995].

References

[Abraham, 2005] Ajith Abraham. Artificial neural networks. *Handbook of measuring system design*, 2005.

[Adadi and Berrada, 2020] Amina Adadi and Mohammed Berrada. Explainable ai for healthcare: From black box to interpretable models. In *Embedded Systems and Artificial Intelligence*, pages 327–337. Springer, 2020.

[Ahn *et al.*, 2016] Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. A neural knowledge language model. *arXiv preprint arXiv:1608.00318*, 2016.

[Anand and Büchner, 1998] Sarabjot S Anand and Alex G Büchner. *Decision support using data mining*. Financial Times Management, 1998.

[Augusto *et al.*, 2017] J Augusto, Asier Aztiria, Dean Kramer, and Unai Alegre. A survey on the evolution of the notion of context-awareness. *Applied Artificial Intelligence*, 31(7-8):613–642, 2017.

[Barredo Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[Barto and Sutton, 1995] Andrew G Barto and Richard S Sutton. Reinforcement learning. *Handbook of brain theory and neural networks*, pages 804–809, 1995.

[Bengio *et al.*, 2013] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[Bengio, 2009] Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[Bérard *et al.*, 2016] Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. Multivec: a multilingual and multilevel representation learning toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference*, 2016.

[Bishop, 2006] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[Brenon *et al.*, 2018] Alexis Brenon, François Portet, and Michel Vacher. Context feature learning through deep learning for adaptive context-aware decision making in the home. In *2018 14th International Conference on Intelligent Environments (IE)*, pages 32–39. IEEE, 2018.

[Brézillon, 1999] Patrick Brézillon. Context in Artificial Intelligence: I. A survey of the literature. *Computers and artificial intelligence*, 18:321–340, 1999.

[Buchanan and Shortliffe, 1984] Bruce G Buchanan and Edward H Shortliffe. Rule-based expert systems: the MYCIN experiments of the stanford heuristic programming project. 1984.

[Burgt, 2020] Joost van der Burgt. Explainable AI in banking. *Journal of Digital Banking*, 4(4):344–350, 2020.

[Cabena *et al.*, 1998] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., 1998.

[Capone and Bertolaso, 2020] Luca Capone and M. Bertolaso. A philosophical approach for a human-centered explainable AI. In *The Italian Workshop on Explainable Artificial Intelligence, the 19th International Conference of the Italian Association for Artificial Intelligence*, 2020.

[Cayton, 2005] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, pages 1–17, 2005.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *The 18th Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Chung *et al.*, 2016] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704*, 2016.

[Cios *et al.*, 2007] Krzysztof J Cios, Witold Pedrycz, Roman W Swiniarski, and Lukasz Andrzej Kurgan. *Data mining: a knowledge discovery approach*. Springer Science & Business Media, 2007.

[Das, 2003] Amit Das. Knowledge representation. In Hossein Bidgoli, editor, *Encyclopedia of Information Systems*, pages 33–41. Elsevier, New York, 2003.

[Davis *et al.*, 1993] Randall Davis, Howard Shrobe, and Peter Szolovits. What is a knowledge representation? *AI magazine*, 14(1):17–17, 1993.

[Davis, 2015] Ernest Davis. Knowledge representation. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, pages 98–104. Elsevier Inc., 2015.

[Dietterich *et al.*, 2008] Thomas G Dietterich, Pedro Domingos, Lise Getoor, Stephen Muggleton, and Prasad Tadepalli. Structured machine learning: the next ten years. *Machine Learning*, 73(1):3, 2008.

[Dolgikh, 2018] Serge Dolgikh. Spontaneous concept learning with deep autoencoder. *International Journal of Computational Intelligence Systems*, 12(1):1, 2018.

[Dong *et al.*, 2020] Yuxiao Dong, Ziniu Hu, Kuansan Wang, Yizhou Sun, and Jie Tang. Heterogeneous network representation learning. In *The 29th International Joint Conference on Artificial Intelligence*, pages 4861–4867, 2020.

[Ehsan *et al.*, 2021] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. Operationalizing

human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.

[Fayyad *et al.*, 1996] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

[Fazi, 2020] M Beatrice Fazi. Beyond human: Deep learning, explainability and representation. *Theory, Culture & Society*, pages 1–23, 2020.

[Fletcher *et al.*, 1991] Peter Fletcher, Hughes Hoyle, and C Wayne Patty. *Foundations of discrete mathematics*. Brooks/Cole, 1991.

[Gao *et al.*, 2019] Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. Explainable recommendation through attentive multi-view learning. *The 33th AAAI Conference on Artificial Intelligence*, 33(01):3622–3629, Jul. 2019.

[Gilpin *et al.*, 2018] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *The 5th IEEE International Conference on data science and advanced analytics*, pages 80–89. IEEE, 2018.

[Goebel *et al.*, 2018] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction*, pages 295–303. Springer, 2018.

[Gordon and Desjardins, 1995] Diana F Gordon and Marie Desjardins. Evaluation and selection of biases in machine learning. *Machine learning*, 20(1-2):5–22, 1995.

[Grazzini and Pantisano, 2015] Jacopo Grazzini and Francesco Pantisano. Collaborative research-grade software for crowd-sourced data exploration: from context to practice - part I: Guidelines for scientific evidence provision for policy support based on big data and open technologies. EUR 27094. *Luxembourg: Publications Office of the European Union*, (JRC94504), 2015.

[Guarino *et al.*, 2009] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

[Gunning, 2017] David Gunning. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*, 2(2), 2017.

[Hagedorn *et al.*, 2020] Thomas Hagedorn, Mary Bone, Benjamin Kruse, Ian Grosse, and Mark Blackburn. Knowledge representation with ontologies and semantic web technologies to promote augmented and artificial intelligence in systems engineering. *Insight*, 23(1):15–20, 2020.

[Hamilton, 2020] William L Hamilton. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, 14(3):1–159, 2020.

[He *et al.*, 2021] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

[Heuillet *et al.*, 2021] Alexandre Heuillet, Fabien Couthouis, and Natalia Diaz-Rodriguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Hollister *et al.*, 2017] Debra L Hollister, Avelino Gonzalez, and James Hollister. Contextual reasoning in human cognition and the implications for artificial intelligence systems. In *The International and Interdisciplinary Conference on Modeling and Using Context*, pages 599–608. Springer, 2017.

[Honavar, 1995] Vasant Honavar. Symbolic artificial intelligence and numeric artificial neural networks: towards a resolution of the dichotomy. In *Computational architectures integrating neural and symbolic processes*, pages 351–388. Springer, 1995.

[Ietswaart *et al.*, 2021] Robert Ietswaart, Benjamin M Gyori, John A Bachman, Peter K Sorger, and L Stirling Churchman. Genewalk identifies relevant gene functions for a biological context using network representation learning. *Genome biology*, 22(1):1–35, 2021.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations*, 2014.

[Le Saux *et al.*, 1999] Elisabeth Le Saux, Philippe Lenca, Philippe Picouet, and Jean-Pierre Barthélemy. An anthropocentric tool for decision making support. In *The International Joint Conference on Artificial Intelligence*, volume 16, pages 338–343. Lawrence Erlbaum Associates LTD, 1999.

[Le Saux *et al.*, 2002] Elisabeth Le Saux, Philippe Lenca, and Philippe Picouet. Dynamic adaptation of rules bases under cognitive constraints. *European Journal of Operational Research*, 136(2):299 – 309, January 2002.

[Leavy, 2018] Susan Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pages 14–16, 2018.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Lee *et al.*, 2016] Chan-Su Lee, Ahmed Elgammal, and Marwan Torki. Learning representations from multiple manifolds. *Pattern Recognition*, 50:74–87, 2016.

[Lenca, 2002] Philippe Lenca. Human centered processes. *European Journal of Operational Research*, 136(2):231–232, 2002.

[Lepri *et al.*, 2018] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2018.

[Lesort *et al.*, 2017] Timothée Lesort, Mathieu Seurin, Xinrui Li, Natalia Díaz-Rodríguez, and David Filliat. Unsupervised state representation learning with robotic priors: a robustness benchmark. *arXiv preprint arXiv:1709.05185*, 2017.

[Lesort *et al.*, 2018] Timothée Lesort, Natalia Díaz-Rodríguez, Jean-François Goudou, and David Filliat. State Representation Learning for Control: An Overview. *Neural Networks*, 108:379–392, December 2018.

[Li *et al.*, 2018] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2018.

[Lim *et al.*, 2009] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128, 2009.

[Lin *et al.*, 2018] Yankai Lin, Xu Han, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Knowledge representation learning: A quantitative review. *CoRR*, abs/1812.10901, 2018.

[Liu *et al.*, 2020] Alexander H Liu, Tao Tu, Hung-yi Lee, and Lin-shan Lee. Towards unsupervised speech recognition and synthesis with quantized speech representation learning. In *The 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7259–7263. IEEE, 2020.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[Madumal *et al.*, 2020] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *The 34th AAAI Conference on Artificial Intelligence*, pages 2493–2500, Apr. 2020.

[Malhotra and Nair, 2015] Meenakshi Malhotra and TR Gopalakrishnan Nair. Evolution of knowledge representation and retrieval techniques. *International Journal of Intelligent Systems and Applications*, 7(7):18–28, 2015.

[McCulloch and Pitts, 1943] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[Mikolov *et al.*, 2013] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

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

[Mueller *et al.*, 2021] Shane T Mueller, Elizabeth S Veinott, Robert R Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J Clancey. Principles of explanation in human-ai systems. *arXiv preprint arXiv:2102.04972*, 2021.

[Nelson, 2019] Gregory S Nelson. Bias in artificial intelligence. *North Carolina medical journal*, 80(4):220–222, 2019.

[Nguyen *et al.*, 2021] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth. In *9th International Conference on Learning Representations, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Puiutta and Veith, 2020] Erika Puiutta and Eric M. S. P. Veith. Explainable reinforcement learning: A survey. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar R. Weippl, editors, *Machine Learning and Knowledge Extraction - 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25-28, 2020, Proceedings*, volume 12279 of *Lecture Notes in Computer Science*, pages 77–95. Springer, 2020.

[Qi *et al.*, 2020] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Towards more explainability: Concept knowledge mining network for event recognition. In *The 28th ACM International Conference on Multimedia*, page 3857–3865, New York, NY, USA, 2020. Association for Computing Machinery.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust you?": Explaining the predictions of any classifier. In *The 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Roberts *et al.*, 2018a] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *The 35th International Conference on Machine Learning*, pages 4364–4373. PMLR, 2018.

[Roberts *et al.*, 2018b] Adam Roberts, Jesse H Engel, Sageev Oore, and Douglas Eck. Learning latent representations of music to generate interactive musical palettes. In *IUI Workshops*, 2018.

[Rojat *et al.*, 2021] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable artificial intelligence (XAI) on Time Series Data: A Survey. *arXiv preprint arXiv:2104.00950*, 2021.

[Rumelhart *et al.*, 1986] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[Sallinger *et al.*, 2020] Emanuel Sallinger, Sahar Vahdati, and Mojtaba Nayyeri. Workshop knowledge representation & representation learning, 2020.

[Schlegel *et al.*, 2019] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, D. Oelke, and D. Keim. Towards a rigorous evaluation of XAI methods on time series. *The 2019 IEEE/CVF International Conference on Computer Vision Workshop*, pages 4197–4201, 2019.

[Shapiro, 2006] Stuart C Shapiro. Knowledge representation. *Encyclopedia of cognitive science*, 2006.

[Shortliffe, 1974] Edward H Shortliffe. A rule-based computer program for advising physicians regarding antimicrobial therapy selection. In *Proceedings of the 1974 annual ACM conference-Volume 2*, pages 739–739, 1974.

[Soleimani *et al.*, 2021] Melika Soleimani, Ali Intezari, Nazim Taskin, and David Pauleen. Cognitive biases in developing biased Artificial Intelligence recruitment system. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 5091, 2021.

[Su *et al.*, 2011] Ya Su, Yun Fu, Xinbo Gao, and Qi Tian. Discriminant learning through multiple principal angles for visual recognition. *IEEE transactions on image processing*, 21(3):1381–1390, 2011.

[Sun *et al.*, 2020] Ke Sun, Lei Wang, Bo Xu, Wenhong Zhao, Shyh Wei Teng, and Feng Xia. Network representation learning: From traditional feature learning to deep learning. *IEEE Access*, 8:205600–205617, 2020.

[Torki *et al.*, 2010] Marwan Torki, Ahmed Elgammal, and Chan Su Lee. Learning a joint manifold representation from multiple data sets. In *The 20th International Conference on Pattern Recognition*, pages 1068–1071. IEEE, 2010.

[Tu *et al.*, 2018] Cunchao Tu, Xiangkai Zeng, Hao Wang, Zhengyan Zhang, Zhiyuan Liu, Maosong Sun, Bo Zhang, and Leyu Lin. A unified framework for community detection and network representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1051–1065, 2018.

[Van Fraassen, 1988] Bas Van Fraassen. The pragmatic theory of explanation. *Theories of Explanation*, 8:135–155, 1988.

[Vilone and Longo, 2020] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[Walton, 2004] Douglas Walton. A new dialectical theory of explanation. *Philosophical Explorations*, 7(1):71–89, 2004.

[Wang *et al.*, 2020] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence (accepted)*, 2020.

[Wirth, 2000] Rüdiger Wirth. Crisp-dm: Towards a standard process model for data mining. In *The 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39, 2000.

[Wu *et al.*, 2020] Lirong Wu, Zicheng Liu, Zelin Zang, Jun Xia, Siyuan Li, and Stan Z Li. Deep clustering and representation learning that preserves geometric structures. *arXiv e-prints*, pages arXiv–2009, 2020.

[Xie *et al.*, 2018] Ruobing Xie, Zhiyuan Liu, Fen Lin, and Leyu Lin. Does William Shakespeare REALLY Write Hamlet? Knowledge Representation Learning With Confidence. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4954–4961. AAAI Press, 2018.

[Xu *et al.*, 2013] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[Xu *et al.*, 2018] Cai Xu, Ziyu Guan, Wei Zhao, Yunfei Niu, Quan Wang, and Zhiheng Wang. Deep multi-view concept learning. In *The 27th International Joint Conference on Artificial Intelligence*, pages 2898–2904, 2018.

[Yang *et al.*, 2015] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. Network representation learning with rich text information. In *The 24th International Joint Conference on Artificial Intelligence*, volume 2015, pages 2111–2117, 2015.

[Yin *et al.*, 2016] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. Neural generative question answering. In *The 25th International Joint Conference on Artificial Intelligence*, 2016.

[Zhang *et al.*, 2011] Zhenyue Zhang, Jing Wang, and Hongyuan Zha. Adaptive manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):253–265, 2011.

[Zhang *et al.*, 2016] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. Collaborative knowledge base embedding for recommender systems. In *The 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362, 2016.

[Zhang *et al.*, 2020] D. Zhang, J. Yin, X. Zhu, and C. Zhang. Network representation learning: A survey. *IEEE Transactions on Big Data*, 6(01):3–28, jan 2020.

[Zhong *et al.*, 2016] Guoqiang Zhong, Li-Na Wang, Xiao Ling, and Junyu Dong. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, 2(4):265–278, 2016.

[Zhu *et al.*, 2014] Zhenfeng Zhu, Linlin Du, Lei Zhang, and Yao Zhao. Shared subspace learning for latent representation of multi-view data. *Journal of Information Hiding and Multimedia Signal Processing*, 5(3):546–554, 2014.

Streaming and Learning the Personal Context

Fausto Giunchiglia[1*], **Marcelo Rodas Britez**[1], **Andrea Bontempelli**[1] and **Xiaoyue Li**[1]
[1]University of Trento, Trento, Italy
{fausto.giunchiglia, marcelo.rodasbritez, andrea.bontempelli, xiaoyue.li}@unitn.it

Abstract

The representation of the personal context is complex and essential to improve the help machines can give to humans for making sense of the world, and the help humans can give to machines to improve their efficiency. We aim to design a novel model representation of the personal context and design a learning process for better integration with machine learning. We aim to implement these elements into a modern system architecture focus in real-life environments. Also, we show how our proposal can improve in specifically related work papers. Finally, we are moving forward with a better personal context representation with an improved model, the implementation of the learning process, and the architectural design of these components.

1 Introduction

Every person makes sense of their personal context differently because of their different sets of personal characteristics (intelligence) and behaviour (life choices). However, the machine's understanding of the personal context is radically different from the user's understanding. This limitation is due to the limited definition of the personal context, and the lack of tools to make sense of the personal context. For instance, while the person you are with now can be linked to a name, for people it has more meaning than just a name, e.g., friend and colleague. Additionally, these meanings are not fixed, they may change at any time, and every person can assign additional meaning using different criteria. Thus, effective context recognition requires a complex and dynamic representation of the personal context and the collaboration of the people to fill the cognitive gap of machines.

The addition of human collaboration into the context recognition learning of machines is an important part of supervised machine learning [Vapnik, 2013]. These interactions bring new challenges to the implementation of machine learning algorithms. For instance, humans can be defined as the expert of the supervised algorithms, interacting in an offline fashion by annotating sensor data [Webb, 2003], or the interaction can be directly online, as active learning [Settles, 2009; Hoque and Stankovic, 2012; Hossain *et al.*, 2017]. The human collaboration is important when we are moving into real-life scenarios [Kwapisz *et al.*, 2011].

Other challenges of this collaborative approach are the possibility of overwhelming the humans and the possible differences between the assignment of meaning between people [Chang *et al.*, 2017], thus, making the annotation a personal activity. Then, having humans as the input of the information opens the possibility of human error in the collaboration process [Tourangeau *et al.*, 2000], and this issue is well known in social sciences and psychology, because of response biases in answering self-reports [West and Sinibaldi, 2013], and more importantly, these biases are not well-understood [Freedman *et al.*, 2013].

We propose a novel context model based on the work from [Giunchiglia *et al.*, 2018]. That work focused on ensuring the reliability of annotations, whereas our focus is on improving personal context representations to get closer to work in real-life scenarios. So, we propose to add a more precise representation of personal context that can also work with machine learning algorithms. We formalize a context model based on ontology and use it with the streaming data to have a knowledge representation of context data. This formalization allows moving towards a generic definition of context that can work with existing multi-label machine learning approaches, using a conversion algorithm. Eventually, the last piece of the puzzle will be the design and development of the Streaming System to manage technically the dynamic context data, and it will be organized in the system architecture with modular components for independent development and easy deployment in current cloud environments.

Some examples of our model improvement can be seen compared with our main related work [Giunchiglia *et al.*, 2018; Bontempelli *et al.*, 2020; Zeni *et al.*, 2019]. All of them can take benefit from our novel personal context representation and can use our conversions algorithms to explicitly implement the transformations needed for the machine learning algorithms.

The paper is structured as follows. Section 2 introduces context modeling. Section 3 illustrates our representation of personal context, while we provide the formal representation in Section 4. Then, we show the learning process to transition from our formal representation to machine learning representation in Section 5 and how our formal representation

*Contact Author

can be converted to a Direct Acyclic Graph (DAG). Finally, Section 6 describes works related to ours, and Section 7 concludes our paper.

2 The context in time

When we talk about the context, we concentrate on the context of a person called observer. The observer's context is the representation of a partial view of the world. We describe this context into three main dimensions: the viewpoint, the part-whole relation, and the endurant-perdurant.

Firstly, we have the viewpoint dimension divided as outside viewpoint and observer viewpoint. The outside viewpoint is the view of an ideal observer who can describe everything from a certain point of view. We distinguish this view from the world's static and dynamic properties. In the static property, there are whatever does not change in time, e.g., mountains, buildings, streets. In the dynamic property, there are not only the moving people, but also the moving animals, and facilities in their manifestations, like trains. Then, the observer viewpoint describes how the observer perceives what is around her or him. In this view, we also have the property of being static or dynamic, but it is relative to the movements of the observer.

Secondly, context is a part-whole relation. In our everyday life, when we do things, we are always embedded in the world. From an ontological view, we are part of the whole world. Thus, we call reference context as the element of the outside context with a volume and extension that is large enough to contain all our movements and changes. For instance, the reference context is the city of Trento when the user walks around, or the users' home when they are at home. In turn, our body as a whole has parts (e.g., arms, mind, legs) that are with us all the time, and they define the internal context of the user. The internal context identifies the elements of the user's body at different levels of abstraction. We usually distinguish between physical parts, such as arms, body, fingers, and mental parts, such as mind, memory, emotions. The context as a part-whole relation is divided into reference context and internal context, and both contexts with different dimensions play a role in our life.

Thirdly, context as endurant and perdurant refers to the bigger relation to changes in time or space. Events and actions are perdurants and elements, like me, are endurants.

2.1 The spatio-temporal context

The context, as viewpoints, defines the reference point from which we construct the context and the context, as part-whole, defines which parts we should consider. So, next, we need to define how we keep track of the context from a quantitative point of view, with a set of quantitative and qualitative measures. Therefore, based on these measures, we introduce the spatio-temporal context.

The spatio-temporal context consists of the temporal and spatial reference context. The former includes dates, times and all the additional notations like weekdays and seasons. The latter contains the world coordinate system. There are various reasons why context should be represented as a spatio-temporal context. First, this is a common representation when we think of the world. Second, any device today can easily retrieve the time and time zones, and the space coordinates (e.g., via GPS). Third, time can be used to measure the changes in all the elements of the world, all evolving at different speeds, thus temporal context allows us to use them together based time. Finally, a lot of data about the spatial reference context and its sub-contexts are available from external sources (e.g., Google Maps, OpenStreetMaps).

The spatio-temporal context, also called objective context, at time t is defined as:

$$\begin{aligned} o_t = (D_t, T_t : L_t,\ me,\ \mathrm{coord}_t(me), \\ P_t^1 : \mathrm{coord}_t(P^1), \ldots, P_t^k : \mathrm{coord}_t(P^k), \\ O_t^1 : \mathrm{coord}_t(O^1), \ldots, O_t^m : \mathrm{coord}_t(O^m)) \end{aligned}$$

where D_n, T_n stands for date and time respectively, L_n is the location, namely the smallest possible spatial reference context that we can compute. Here me is the observer, P^i are persons and O^m are objects. The function $\mathrm{coord}(\ldots)$ computes the spatial coordinates of me, objects, and persons.

The number and type of persons and objects change over time. Hence, we will have a sequence of time-tagged states, namely $\mathcal{O} = \{o_1, \ldots, o_n\}$. We call the sequence $\mathcal{O}$ as the streaming context. In the streaming context, within the given reference location, it is easy to compute spatial relations (e.g., near, right, left, in front, far relative to the location) of the different elements among themselves. For instance, the system can compute that the smartphone is *in* the home building and the smartphone is *near* the computer.

2.2 The objective and subjective context

The spatio-temporal context is also called objective context, since all the relations are computed concerning what is objectively measured, in terms of spatial relations. However, notice that different observers will have different views of the world. For instance, the school building has the function of study-place from the point of view of a student and is the teacher's workplace. The word function here is used with the precise meaning defined in [Giunchiglia and Fumagalli, 2017; Giunchiglia *et al.*, 2018]. Hence, "*the function of an object formalizes the behavior that an object is expected to have*" [Giunchiglia and Fumagalli, 2017]. For instance, objects are trains and buildings. The expected behavior may be the purpose of the object (e.g., fridge) or the role of a person (e.g., friend).

The subjective context includes both the objective context elements and the function of persons and objects as seen by the user. Thus, the subjective context at time t is defined as:

$$\begin{aligned} s_t = &(D_t, T_t :\ L_t, me, \mathrm{coord}_t(me), F_t(P^1), \ldots, F_t(P^k), \\ &F_t(O^1), \ldots, F_t(O^m)), \end{aligned}$$

where $F_n(P^k)$ and $F_n(O^m)$ are the functions with respect to a person P^k and an object O^m, respectively. The number and type of persons and objects, and their functions change over time. The sequence of subjective contexts over time is defined as the subjective streaming context $\mathcal{S} = \{s_1, \ldots, s_n\}$.

2.3 The endurant and perdurant context

In the endurant context [Giunchiglia *et al.*, 2017], its parts are endurants, essentially objects where the spatial extension

of their actions is contained by the space defined by the spatial (reference) context. The actions "*represents how objects change in time*" [Giunchiglia and Fumagalli, 2017]. For instance, running in a park performed by a runner. We also need to represent actions, in particular, the actions that are executed by the endurant me and also by any other elements of the outside dynamic context. Actions can be seen in two ways: *(i)* actions modeled as processes, namely as sequences of single micro-steps, each of length close to zero; *(ii)* actions modeled as events, which are often also called perdurants, namely as complete movements which last for a certain duration. Actions as events have key properties, similar to endurants. An event and an action can associate with a set of component sub-events and sub-actions.

Considering the mentioned concepts, the fundamentally different role of space and time should become clear. Whereas the parts of the space context are only used to limit the space where things happen, the parts of time have the main goal to detail how actions get executed. Things get complicated by adding objects, people, and functions as shown in the data representations shown in Table 1.

It is worth noting that each function of a person is associated with a limited set of actions, and the type of action that a person can perform can be considerably limited by knowing his or her function.

The actions apply to me and persons, whereas functions apply to me, other objects and persons. Notice also, that the stored location L is limited by the most specific location that we can compute. This is because the bigger locations are assumed to be static and stored in the system. For events, instead, we store the smallest possible most general event as well as those component actions which are done during a certain period. Thus, for instance, the action/event meeting can have sub-actions such as talking, walking, listening, typing. Table 2 reports the streaming context matrix of Example 1 and shows how the context changes over time.

3 The current context

The streaming context describes the contexts of an observer according to time. To represent each context occurrence, we define a set of notions to build a figure of the current context. We mainly divide the current context into four types of context according to how things compose in space and time. From top to bottom we have the following cases:

- 1L1E (One Location One Event), such as a lecture holds in a classroom;
- 1LME (One Location Multiple Events), such as a sequence of meetings hold in an office, or eat breakfast, lunch and dinner at home;
- 1EMC (One Event Multiple Locations), such as a travel goes from many different places;
- MEMC (Multiple Events Multiple Locations), it is the most complex case, which mixes with the former three cases.

An 1EMC example is shown in Figure 1, which describes the following travel scenario around observer me.

Example 1 *In the travel scenario, me is Xiaoyue, she has travel named Travel 1 in the Trentino of Italy from 12:00 to 13:00 of 2th, June 2021. From 12:00 to 12:30 on this day, she takes train 1 from Rovereto to Trento, she sits on seat 1 by herself. From 12:30 to 12:55 on the same day, she walks on Roads 2 from Trento Train Station to Xiaoyue's Home, together with her friend named Haonan. In addition, Xiaoyue talks to Haonan, and Haonan listens to Xiaoyue when they are walking. This scenario involves one event travel and multiple locations.*

In general, in the Figure 1, all elements are divided into Perdurants and Endurants. Perdurants are the Event and Action, and Endurants include Person, Object, and Location. An Event happens in a Location, a Person and Object appear in an Event and Action is in a Person. Most elements' inclusion relationships IN can be represented by the positions of those elements' internal boxes. For the top-level Location and Event, we can add an extra attribute box for them respectively as $InLocation(Location)$ and $InEvent(Event)$, to represent their belongs.

In the rest of this section, we list the attributes of Location, Event, Object, Person and Action, each kind of attributes is represented as a box in the figure. The **Location** has the following attributes:

- Spatial properties: Coordinates (x_i, y_i, z_i), Volume (Δx_i, Δy_i, Δy_i) , and InLocation(L_i) shows the super Location for the top level Location;
- Visual properties, namely some properties of Location that can be observed visually;
- Location' functions: $FunctionOf(U)$ with $U \in \{P^1, \ldots, P^k, O^1, \ldots, O^M\}$, which shows location's functions for persons and objects;
- An extra box: including the rest part of the context that happens in the Location.

The **Event** orders by time and is represented by a box with round corners. Events include actual events and virtual events. An Event can have Sub-Events, the Event and Sub-Event have the following attributes:

- Super Event for the top level event: $InEvent(E_i)$, which shows the super event E_i for the current Event;
- Temporal properties: Begin Time - End Time ($Date_i$, $Time_i$ - $Date_j$, $Time_j$), which shows the date and time of the start and end of the event;
- An extra box: including the rest part of the context that happens in the Event.

The **Object** appears in Event and has the following attributes:

- Spatial properties: Coordinates (x_i, y_i, z_i), In/Far/... (P^i/O^m/...);
- Visual properties, namely some properties of Object that can be observed visually;
- Object's functions: $FunctionOf(U)$ with $U \in \{P^1, \ldots, P^k, O^1, \ldots, O^M\}$, which shows Object's functions for persons and other objects.

The **Person** appears in Event and has the following attributes:

$$
\begin{aligned}
&\{\\
&\quad \Big(D_1, T_1 : \mathrm{super}(L_1), \mathrm{super}(E_1), L_1, E_1, me, \mathrm{coord}_1(me), A_1^{me}, F_1(P^1) : A_1^{P^1}, \ldots, F_1(P^k) : A_1^{P^k}, F_1(O^1), \ldots, F_1(O^m)\Big),\\
&\quad \Big(D_2, T_2 : \mathrm{super}(L_2), \mathrm{super}(E_2), L_2, E_2, me, \mathrm{coord}_2(me), A_2^{me}, F_2(P^1) : A_2^{P^1}, \ldots, F_2(P^k) : A_2^{P^k}, F_2(O^1), \ldots, F_2(O^m)\Big),\\
&\quad \ldots,\\
&\quad \Big(D_n, T_n : \mathrm{super}(L_n), \mathrm{super}(E_n), L_n, E_n, me, \mathrm{coord}_n(me), A_n^{me}, F_n(P^1) : A_n^{P^1}, \ldots, F_n(P^k) : A_n^{P^k}, F_n(O^1), \ldots, F_n(O^m)\Big)\\
&\}
\end{aligned}
$$

Table 1: The personal streaming context, where E_n is an event, $\mathrm{super}(L_n)$ and $\mathrm{super}(E_n)$ are the super-classes of L_n or E_n, respectively. The set of actions performed by *me* or by the persons based on their functions are denoted with $A_n^k = \{a_1, \ldots, a_i\}$, with $k \in \{me, P^1, \ldots, P^k\}$.

D_n	T_n	$\mathrm{super}(L_n)$	$\mathrm{super}(E_n)$	L_n	E_n	$\mathrm{coord}_n(me)$	A_n^{me}	$F_n(P^1) : A_n^{P^1}$	$F_n(O^1)$
02/06/2021	12:15	Trentino	Travel 1	Train 1	Take Train	x41, y41, z41	Sitting	*NaN*	RestToolOf(Xiaoyue, Seat 1)
02/06/2021	12:30	Trentino	Travel 1	Roads 2	Walk	x43, y43, z43	Walking, Talking	FriendOf(Xiaoyue, Haonan): Walking, Listening	*NaN*

Table 2: A streaming context matrix representing the travel scenario of Example 1 from the point of view of Xiaoyue, i.e., the observer me. P^1 is Haonan and O^1 is the object "Seat 1". Each column is a property, and every row stands for the current context in a specific timestamp.

- Spatial properties: Coordinates (x_i, y_i, z_i), In/Far/...(P^i/O^m/...);
- Visual properties, namely some properties of Person that can be observed visually;
- Person's functions: $FunctionOf(U)$ with $U \in \{P^1, \ldots, P^k, O^1, \ldots, O^M\}$, which shows Person's functions for other persons and objects;
- Internal states: Physical states ($InPain()$), Mental states ($InMood()$, $InStress()$);
- An extra box: including the Actions of Person.

The **Action** is similar with Event, it orders by time and is represented by a box with round corners. An Action can has many Sub-Action, the Action and Sub-Action have following attributes, each attribute has a box for itself.

- Temporal properties: Begin Time - End Time ($Date_i$, $Time_i$ - $Date_j$, $Time_j$), which shows the date and time of the start and end of the Action;
- Visual properties, namely some properties of Action that can be observed visually;
- Means of the Action: Means(O^m/...);
- Sub-action: Action i;
- Action's functions: $FunctionOf(U)$ with $U \in \{P^1, \ldots, P^k, O^1, \ldots, O^M\}$, which shows Action's functions for persons and objects.

4 The current context as a Knowledge Graph

We use the *Entity Type Graph* (ETG) and the *Entity Graph* (EG) in ontology to represent the context. ETG is a knowledge graph where nodes are entity types, which are further decorated with data properties. The object properties are presented in the graph representing the relations among the entity types.

In Figure 2, white box nodes are entity types that include data properties with data types, the green box nodes enumerates the values of data property. The object properties are connecting all entity types, represented by diamond nodes with arrows. The inheritance relation in the ETG is represented by an arrow from the super-class to the sub-class, and the sub-class inherits all the data properties and object properties of its super-class. The EG populates the entity types and properties defined in the ETG with specific values. It is a data graph where nodes are entities that are connected by object property values representing the relations. Each entity further includes data property values. The streaming context can be viewed as a stream of EGs, in which each EG describes the context at a different time.

We design an EG example as Figure 3 according to the scenario in Example 1. The graph represents the context around "Me", this contains entities shown by nodes, e.g., "Smartphone", "Talking", "Walk". Also, we can see object property values.

5 Learning Context

AI applications like smart personal assistants provide a service to the users based on their context. The context information is usually not available to the machine, and hence it has to infer the location or the activity of the user from sensor data (e.g., GPS, accelerometer, nearby Bluetooth devices). In our scenario, Xiaoyue is carrying a smartphone that generates a stream of sensor readings, and she annotates the data by answering questions about her context, e.g., "Where are you"

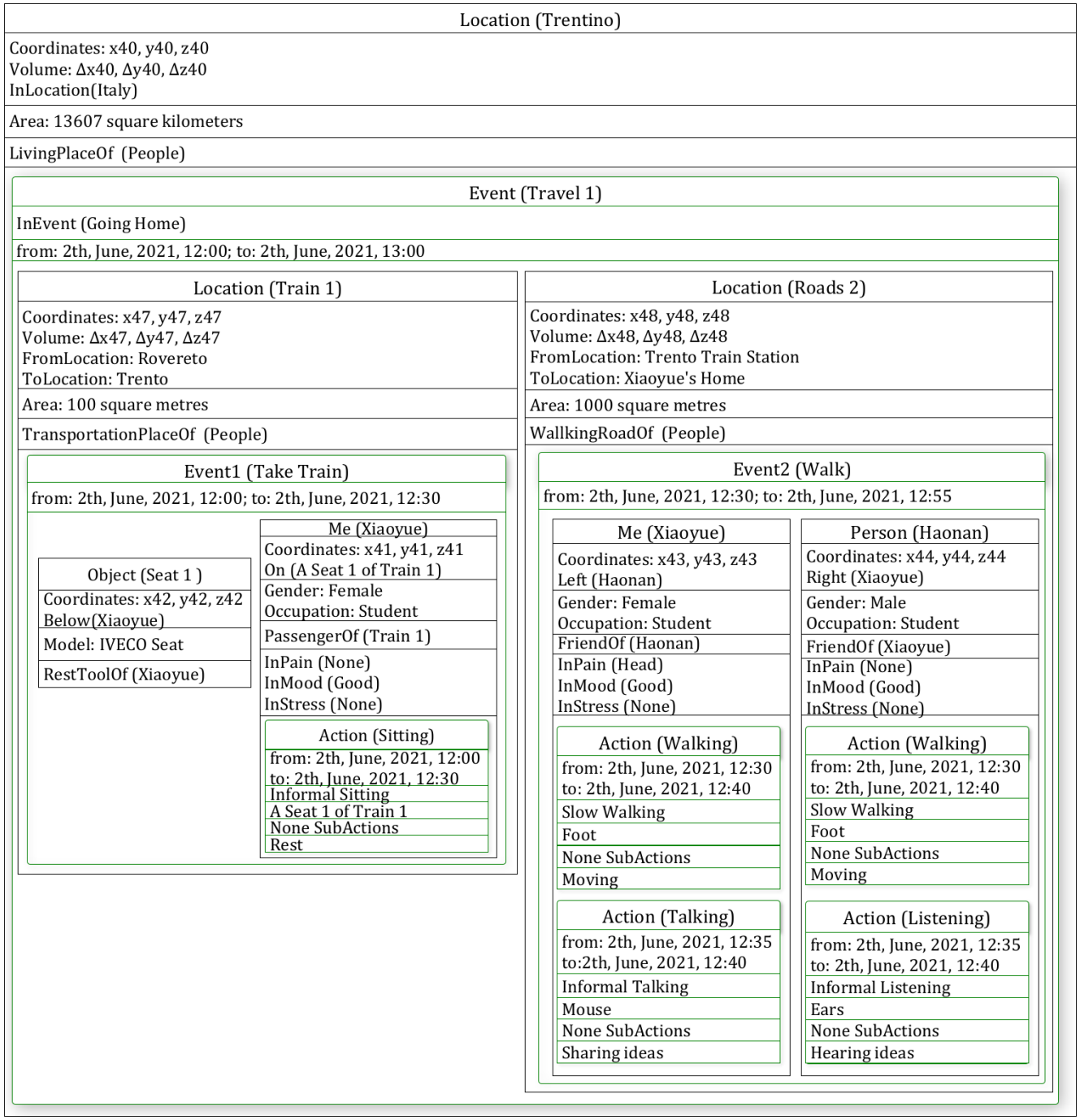


Figure 1: One Event Multiple Locations: a travel around me. Representation of the Example 1.

and "What are you doing?". The sensor data are aggregated in time windows generating a stream of instances (e.g., the average number of nearby Bluetooth devices in the last 30 minutes). On each incoming instance, the machine decides whether to query the user to acquire the labels. The machine learning technique defines the query strategy, and in the simplest case, the labels are acquired on every instance.

The user's context recognition is a supervised learning problem in which an instance $\mathbf{x}$ is associated to multiple concepts $\mathbf{y}$ (aka classes in machine learning). The concepts are organized in a ground-truth hierarchy $\mathcal{H} = (C, I)$, which is a direct acyclic graph (DAG) where nodes $C = \{1, \ldots, c\}$ are the concepts and edges $I \subset C \times C$ are *is-a* relations, i.e., $I = \{(c_i, c_j) \,|\, c_i, c_j \in C \text{ and } c_i \text{ is a child of } c_j\}$ [Silla and Freitas, 2011]. The labels of the instances are indicator vectors $\mathbf{y} \in \{0, 1\}^c$, where the i-th elements is 1 if $\mathbf{x}$

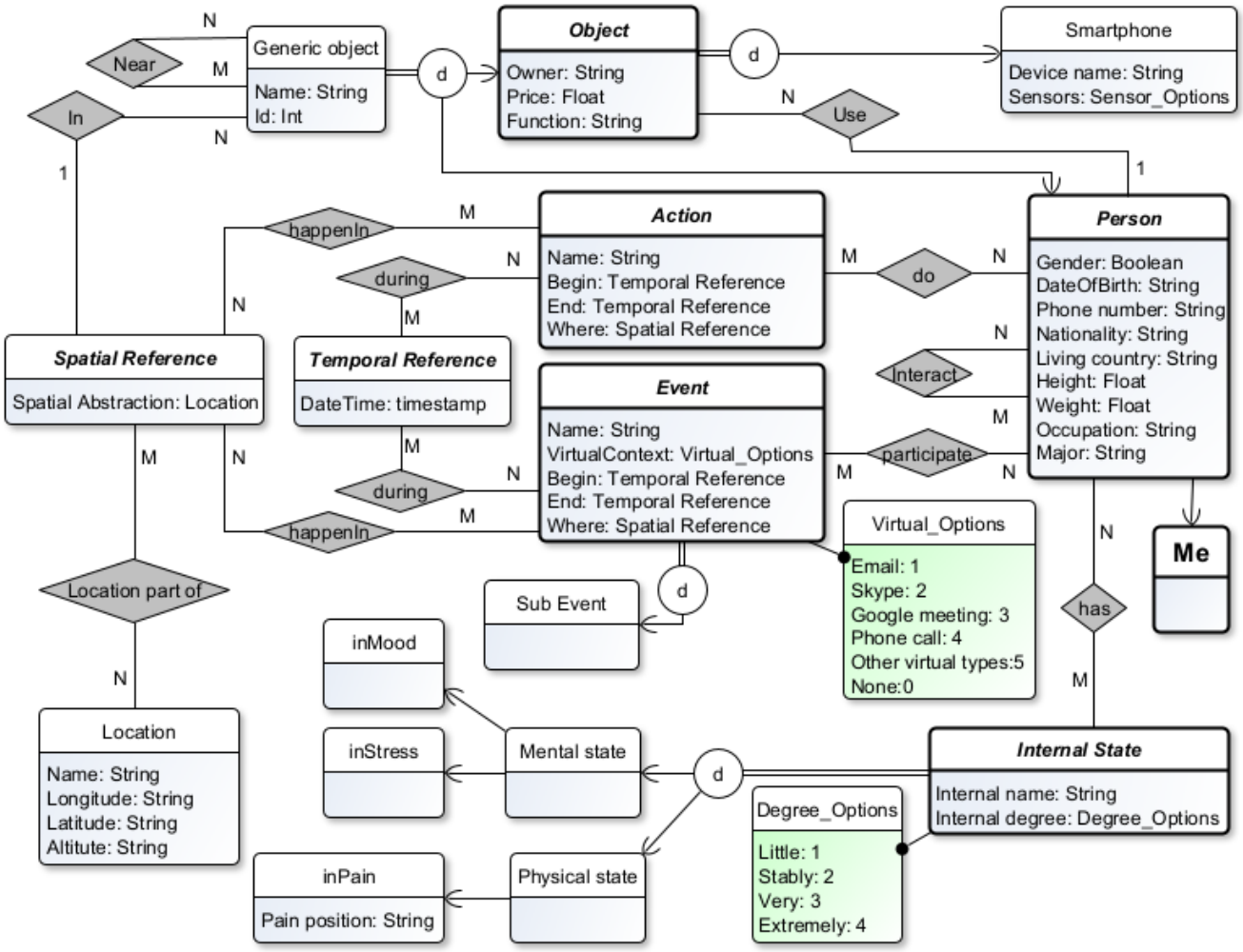


Figure 2: An ETG representing partially the personal context in our travel example.

belong to i-th concept in $\mathcal{H}$ and 0 otherwise. The machine is trained on a stream of examples $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{y}_t)$ drawn from a ground-truth distribution $P(\mathbf{X}, \mathbf{Y})$ that is always consistent with a ground-truth hierarchy $\mathcal{H}$, i.e., if there is an edge from class c_i to class c_j, then $y^i = 1$ implies $y^j = 1$ and conversely $y^j = 0$ implies $y^i = 0$. The goal of this hierarchical classification tasks is to learn a classifier that recognize well the context on future sensor readings.

The ETG and EG introduced in Section 4 can be used as prior knowledge about the structure of the hierarchy. They encode the available information about the user and the world. Algorithm 1 shows the conversion from ETG and EG to a DAG $\mathcal{H}$. The first step is to convert each entity type (etype) in the ETG as a node in $\mathcal{H}$ (lines 3 - 5). Second, each entity in EG also becomes a node that is added as a child of the node referring to the etype of the entity (lines 6 - 10). The hierarchy encodes the information about the current user, so the *Me* etype and the corresponding entity (e.g., Xiaoyue entity in Figure 3) are not considered.

The properties of the ETG are grouped in properties that are context depends and properties that are static. The value of the former changes every time the users change their context and, in Figure 2 are $Q = \{$*near*, *use*, *interact*, *in*, *do*, *happenIn*, *during*, *participate*$\}$. For instance, if Xiaoyue travel from the city of Trento to Rovereto, the *in* property will change accordingly. In contrast, the fact that Trento is *partOf* Italy can be assumed to remain valid even if user's context is changed. This distinction is necessary since the value of context-dependent properties are derived from the output of the context recognition task (e.g., the machine recognizes that the user is in the city of Trento and thus updates the *in* property accordingly). The object properties that are not contextual are converted as follow. Given a object property $p \in \{$*isA*, *partOf*, *has*$\}$ that links the etype A to B, then the node referring to etype A becomes a child of the node of etype B. For the other object properties, a new node referring to the property is added as child of the codomain etype node (lines 13 - 19). For every object property value i of the prop-

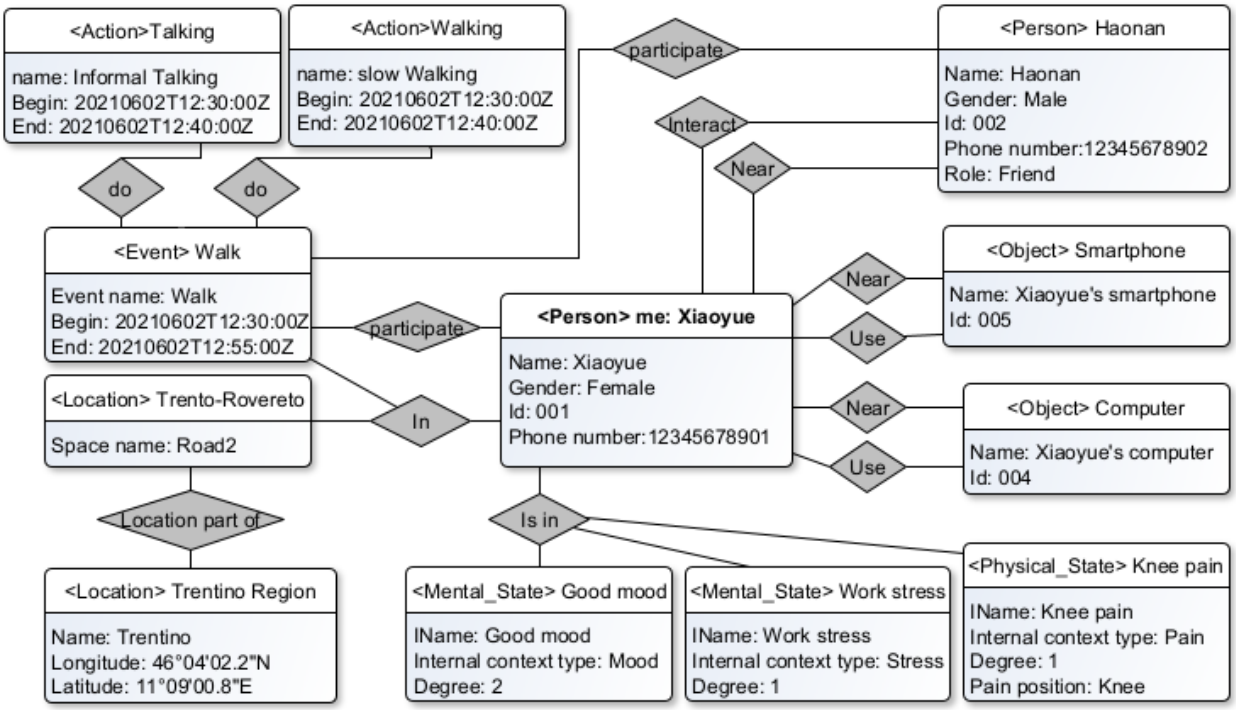


Figure 3: An EG representing partially the scenario about travelling of Example 1.

erty p linking the entities a and b, a new node c_i is added as child of c_p (i.e., the node referring to the property P), and as parent of c_b (i.e., the node pointing to the entity b) (lines 20 - 25). Finally, all nodes that does not have a parent are connect to the root node and the transitive reduction is applied (lines 27 - 30). Figure 4 shows an extract of the DAG resulting from applying Algorithm 1 on EG and ETG presented in Section 4.

Every node in $\mathcal{H}$ has a unique identifier that is used to reference back to the ETG and EG. The node name can be translated into a human-readable text that is used to interact with the user. This aspect is left as future work. The concept hierarchy available at the beginning can evolve over time and has to be continually updated. This aspect has been defined as knowledge drift and is addressed in [Bontempelli *et al.*, 2021].

6 Related Work

Considering our novel context modelling of the personal context, the context learning, and the architecture presented in this paper, we can show as use cases the papers that individually compartmentalizes examples for these improvements. Current works on context recognition focus on learning the relationship between the input data (sensor readings) and the target concepts (context). The structure of the context is implicitly learned by the implemented machine learning algorithm during the training phase. The new parts are described in Section 3 and compared with related work we can outline the following: 1) the connection between the learned context model; 2) The extension of new dimensions

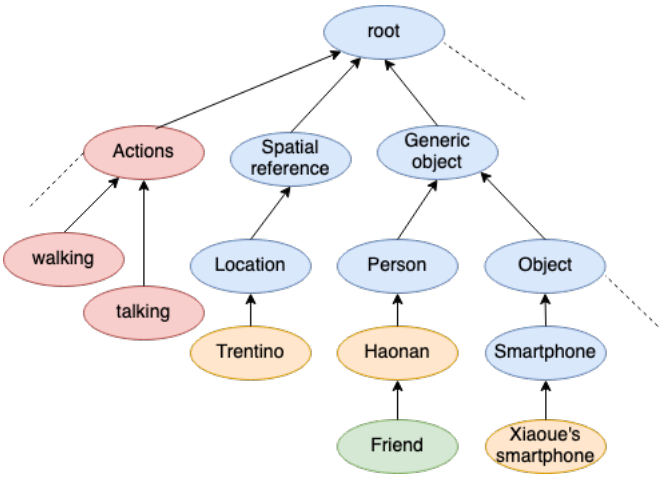


Figure 4: A partial representation of the DAG obtained from ETG and EG examples in Section 4. Orange nodes are derived from EG. Blue nodes are derived from ETG. Red nodes are actions and green nodes are the functions.

Algorithm 1 Convert ETG and EG in a DAG.

Inputs: ETG, EG, and the set of context dependent object properties Q
Outputs: $\mathcal{H} = (C, I)$

$C \leftarrow \emptyset$
$I \leftarrow \emptyset$
for every etype $A \neq$ *Me* in ETG **do**
  let c_A being the node of etype A
  $C = C \cup \{c_A\}$

for every entity a such that $\neg$ Me(a) in EG **do**
  let c_a being the node of the entity a
  let c_A being the node of the etype of a
  $C = C \cup \{c_a\}$
  $I = I \cup \{(c_a, c_A)\}$

for every object property p in ETG such that $p \notin \mathcal{Q}$ **do**
  let c_p being the node of p
  for every $p(A, B)$ in ETG **do**
    let c_A and c_B being the nodes of entity A and B respectively
    if $p \in \{$*isA*, *partOf*, *has*$\}$ **then**
      $I = I \cup \{(c_A, c_B)\}$
    else
      $C = C \cup \{c_p\}$
      $I = I \cup \{(c_p, c_B)\}$
  for every $p(a, b)$ in EG **do**
    let c_i being the node encoding $p(a, b)$
    let c_b being the node of entity b
    $C = C \cup \{c_i\}$
    $I = I \cup \{(c_i, c_p)\}$
    $I = I \cup \{(c_b, c_i)\}$

let c_0 being the root node
$C = C \cup \{c_0\}$
for every node $c \in C$ that $\text{parent}(c) = \emptyset$ **do**
  $I = I \cup \{(c, c_0)\}$ add an edge from c to c_0
apply transitive reduction on $\mathcal{H} = (C, I)$

and classification of context, i.e. Internal State, Functions, Actions. The Algorithm 1 bridges the gap between the two modellings allowing context recognition to leverage both machine learning solutions and knowledge graph representation (ETG and EG). For instance, the ETG and EG can generate the questions needed by the machine to interact with the users. For instance, the personal context recognition model shown in [Giunchiglia *et al.*, 2018], showed to be a good approach to increase the accuracy of the context recognition algorithms. Our work described in Section 3 enhance the representation of the personal context with the hope to perform better in real-life scenarios. In practice, existing approaches for context recognition using batch or streaming sensor data [Vaizman *et al.*, 2018; Bontempelli *et al.*, 2020; Zeni *et al.*, 2019] do not leverage on an explicit context modeling. The context representation is implicit in the labels used to train the machine learning model. The context formalization introduced in this work can be used to structure the output of these machine learning models according to our representation. Moreover, it can help machine learning approaches to interact with users. For instance, the machine can ask if Haonan is a friend of Xiaoyue since they are walking together. Approaches that use active learning strategies (e.g., [Settles, 2009; Hoque and Stankovic, 2012; Hossain *et al.*, 2017]) can benefit of our representation.

Also, existing frameworks for creating context-aware mobile applications, such as Ferreira *et al.*, do not consider the modelling of the context.

7 Conclusion and Future Work

In this paper, we moved forward with a better representation of personal context in real-life environments. We proposed an improved representation of the personal context, adding the internal state, functions, and actions. The learning aspect of our work is the formal definition of an algorithm to transform the streaming input data to ML algorithms. We will put all these components in the system architecture.

In comparison with the work on personal context recognition for human-machine collaboration [Giunchiglia *et al.*, 2018], we have shown an enhancement related to the model representation of the personal context.

Additionally, we have shown how our novel personal context representation can also be leveraged by machine learning algorithms to include prior knowledge about the structure of their output and can be used to drive the interaction with the user. Future work will focus on evaluating the impact of our formalization on an existing approach for fixing mislabeled data when learning the users' contexts [Zeni *et al.*, 2019].

The next step is to propose and implement a modern design of the services related to iLog [Zeni *et al.*, 2014] by a centralised streaming system and linking the personal context data collections with other distributed services of machine learning. This implementation will allow us to test and evaluate our novel context model in near real-life scenarios.

Acknowledgements

The research conducted by Fausto and Xiaoyue has received funding from the European Union's Horizon 2020 FET Proactive project "WeNet – The Internet of us", grant agreement No 823783.

The research conducted by Marcelo and Andrea has received funding from the *"DELPhi - DiscovEring Life Patterns"* project funded by the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) 2017 – DD n. 1062 del 31.05.2019.

References

[Bontempelli *et al.*, 2020] Andrea Bontempelli, Stefano Teso, Fausto Giunchiglia, and Andrea Passerini. Learning in the wild with incremental skeptical gaussian processes. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2886–2892, 7 2020.

[Bontempelli *et al.*, 2021] Andrea Bontempelli, Fausto Giunchiglia, Andrea Passerini, and Stefano Teso. Human-in-the-loop handling of knowledge drift. *arXiv preprint arXiv:2103.14874*, 2021.

[Chang *et al.*, 2017] Yung-Ju Chang, Gaurav Paruthi, Hsin-Ying Wu, Hsin-Yu Lin, and Mark W Newman. An investigation of using mobile and situated crowdsourcing to collect annotated travel activity data in real-word settings. *International Journal of Human-Computer Studies*, 102:81–102, 2017.

[Ferreira *et al.*, 2015] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[Freedman *et al.*, 2013] Vicki A. Freedman, Jessica Broome, Frederick Conrad, and Jennifer C. Cornman. Interviewer and respondent interactions and quality assessments in a time diary study. *Electronic international journal of time use research*, 10(1):55, 2013.

[Giunchiglia and Fumagalli, 2017] Fausto Giunchiglia and Mattia Fumagalli. Teleologies: Objects, actions and functions. In *International conference on conceptual modeling*, pages 520–534. Springer, 2017.

[Giunchiglia *et al.*, 2017] Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. Personal context modelling and annotation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 117–122. IEEE, 2017.

[Giunchiglia *et al.*, 2018] Fausto Giunchiglia, Mattia Zeni, and Enrico Big. Personal context recognition via reliable human-machine collaboration. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 379–384. IEEE, 2018.

[Hoque and Stankovic, 2012] Enamul Hoque and John Stankovic. Aalo: Activity recognition in smart homes using active learning in the presence of overlapped activities. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 139–146. IEEE, 2012.

[Hossain *et al.*, 2017] HM Sajjad Hossain, Md Abdullah Al Hafiz Khan, and Nirmalya Roy. Active learning enabled activity recognition. *Pervasive and Mobile Computing*, 38:312–330, 2017.

[Kwapisz *et al.*, 2011] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[Settles, 2009] Burr Settles. Active learning literature survey. 2009.

[Silla and Freitas, 2011] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 2011.

[Tourangeau *et al.*, 2000] Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. The psychology of survey response. 2000.

[Vaizman *et al.*, 2018] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. 1(4), January 2018.

[Vapnik, 2013] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[Webb, 2003] Andrew R. Webb. *Statistical pattern recognition*. John Wiley & Sons, 2003.

[West and Sinibaldi, 2013] Brady T. West and Jennifer Sinibaldi. The quality of paradata: A literature review. *Improving surveys with paradata*, pages 339–359, 2013.

[Zeni *et al.*, 2014] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. Multi-device activity logging. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous Computing: Adjunct publication*, pages 299–302, 2014.

[Zeni *et al.*, 2019] Mattia Zeni, Wanyi Zhang, Enrico Bignotti, Andrea Passerini, and Fausto Giunchiglia. Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):1–23, 2019.

People in the Context – an Analysis of Game-based Experimental Protocol

Krzysztof Kutt[1*], **Laura Żuchowska**[2], **Szymon Bobek**[1,2] and **Grzegorz J. Nalepa**[1,2]

[1]Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI) and Institute of Applied Computer Science, Jagiellonian University, Kraków, Poland

[2]Department of Applied Computer Science, AGH University of Science and Technology, Kraków, Poland

krzysztof.kutt@uj.edu.pl, szymon.bobek@uj.edu.pl, gjn@gjn.re

Abstract

The paper provides insights into two main threads of analysis of the BIRAFFE2 dataset concerning the associations between personality and physiological signals and concerning the game logs' generation and processing. Alongside the presentation of results, we propose the generation of event-marked maps as an important step in the exploratory analysis of game data. The paper concludes with a set of guidelines for using games as a context-rich experimental environment.

1 Introduction and Motivation

The development of a good personalised intelligent assistant that behaves in a natural way requires the development of proper toolbox as a base [Nalepa *et al.*, 2019]. In order to be user-friendly, an assistant should not only perform its task, but also respond to the user's changing emotions. This is due to our natural tendency to anthropomorphize interfaces – the user will assume that the assistant will react appropriately, e.g., understand that the nervousness is due to a mistake committed. Such affective information can be extracted from the range of physiological signals, particularly obtained through low-cost wearable devices that will make this technology available to everyone. Finally, it is important to note that emotions do not happen in a void—they are always dependent on the context a person is in [Prinz, 2006]—so it is also important to collect information about the user's current situation (e.g., activity, weather conditions, time of day).

An important step in establishing the above-outlined framework for personalized assistants is the collection of the right data. This, in turn, strictly depends on the development of appropriate research environments and experimental protocols. Such issues are addressed in the BIRAFFE (*Bio-Reactions and Faces for Emotion-based Personalization*) series of experiments [Kutt *et al.*, 2021]. Their main objective is to develop methods for emotion recognition using a range of contextual information and physiological signals such as cardiac activity (ECG), electrodermal response (EDA), hand movements (accelerometer) or changes in facial expression. In order to ensure that the research is highly ecological in measurement and easily extendable to wider research groups, wearable and portable, affordable-for-all devices are used.

A key aspect of the BIRAFFE experiments is the use of games as the experimental environment. They were chosen as a trade-off between a stimulus-rich complex near-natural environment and the need to control and record as much context as possible to provide the most detailed post-experimental analyses. The latest version of the experiment, BIRAFFE2 [Kutt *et al.*, 2020][1], used a game consisting of three independent levels. The aim of the first was to evoke positive emotions. The second was intended to induce irritation and frustration, e.g., through impaired control. Finally, the third level was a neutral maze. A detailed description of the games is presented in [Żuchowska *et al.*, 2020].

This paper provides insights into the core analyses of the BIRAFFE2 dataset on contextual information processing in affective games. The first thread, presented in Sect. 2, focuses on the analysis of the relationship between physiological signals and the so-called "Big Five" personality traits. The existence of such relationships in the data will allow further work to create emotion prediction models that will be moderated and personalised through the identification of personality profiles. The second topic, described in Sect. 3, addresses the topic of accurate game logging and the possibility of reconstructing an entire game from such stored logs. The whole article concludes with a set of lessons-learned regarding the implementation of games as an experimental environment in Sect. 4.

2 Physiological Signals and Personality

Before undertaking the analyses, three features were calculated for ECG signal using HeartPy library [van Gent *et al.*, 2019]: heart rate (number of heart beats per minute), mean of successive differences between R-R intervals (MoSD) and breathing rate. Also, to group the valence and arousal scores into discrete variable, 16 clusters were introduced as presented on Fig. 1.

In order to find correlations and dependencies between physiological data (the ECG signal was chosen as an illustration) and personality traits (each on $[1, 10]$ scale), several

*Corresponding Author

[1]The entire dataset from the BIRAFFE2 experiment is available under CC licence on the Zenodo platform, DOI:10.5281/zenodo.3865859.

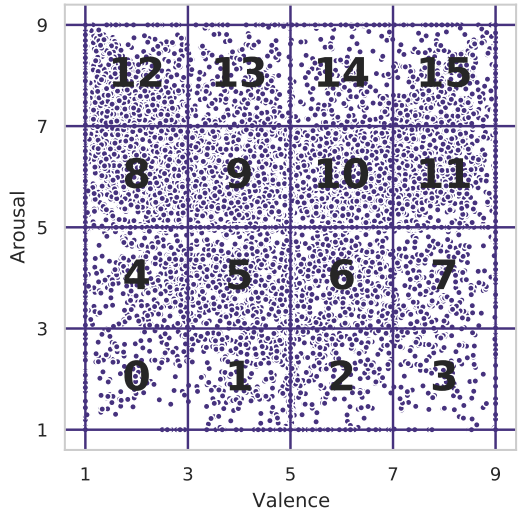


Figure 1: Valence and arousal scores grouped into clusters.

Personality Trait	Mean	SD	Median
Conscientiousness	5.68	2.61	6
Openness	5.48	2.22	6
Agreeableness	6.25	2.39	6
Extraversion	5.84	2.29	7
Neuroticism	5.38	2.45	5

Table 1: Descriptive statistics for personality traits.

approaches to statistical analysis were made. Firstly, basic descriptive statistics were calculated to find outliers and possible extremas. As can be seen in Tab. 1-2, the data was distributed proportionally in terms of mean, median and standard deviation, which indicates a promising start for further analysis.

The second analysis was aimed at investigation of correlations between features. Although the results did not show any strong dependencies between them (see Fig. 2), they indicated the existence of potentially interesting relationships worthy of further analysis and further research. Namely, in terms of the associations between personality and widget responses, valence and arousal are related to distinct traits. For arousal, the highest values are for openness and conscientiousness. On the other hand, valence's most significant factors are agreeableness and extraversion. When considering the correlations between physiological reactions and widget, among heart rate, MoSD, and respiratory rate, the highest values were noted for the first of these for both valence and arousal. The outcome of personality trait to heart rate was presented as maximal for both conscientiousness and extraversion. Considering the MoSD, highest value—and the highest inter-correlation in general, i.e., the correlation between different data sources—was for extraversion (0.23) and conscientiousness (-0.19). Finally, values of correlation for breathing rate played in favor of extraversion.

The last statistical analysis performed was two ANOVAs for valence and arousal (see Tab. 3-4), which indicated several strong associations. What seems most interesting is the strong relationship between heart rate and valence, which is somehow in opposition to most approaches in which heart rate is used to predict arousal, while other signals such as EDA are mostly used for valence [Dzedzickis *et al.*, 2020].

ECG characteristic	Mean	SD
Heart rate [BPM]	80.92	16.21
MoSD [ms]	34.41	37.89
Breathing rate [Hz]	0.10	0.12

Table 2: Descriptive statistics for ECG characteristics.

Independent var.	df	MS	F	p
Conscientiousness	1	3.85	0.60	0.44
Openness	1	5.24	0.82	0.37
Agreeableness	1	68.57	10.70	**0.001**
Extraversion	1	6.12	0.95	0.33
Neuroticism	1	0.02	0.004	0.95
Heart rate	1	59.97	9.35	**0.002**
MoSD	1	0.87	0.14	0.71
Breathing rate	1	2.91	0.45	0.50
Residual Error	10881	6.41		

Table 3: ANOVA model for *valence* with five personality traits and three ECG-related characteristics as independent variables.

3 Game Logs and Questionnaires

As noted in the introduction, one motivation for using games as an experimental environment is the ability to frequently sample and log the entire player context. Properly prepared logs should allow the reconstruction of both the level map (the same for each subject) and the course of the entire game for each player. Indeed, this is possible for the games studied. As part of the log analyses, a number of maps were generated, which were verified by comparison with the games and recorded screencasts of the gameplay. These maps can also be used for aggregated analyses, e.g., by plotting all events of one type followed by an initial visual inspection. Fig. 3 shows all the death locations of the protagonist in the first level. One can notice a very high number of deaths in the central room – this is consistent with the observations made during the experiment: this is the first room where players are just getting familiar with the game interface.

Another part of the analysis was the examination of answers from Game Experience Questionnaire [IJsselsteijn *et al.*, 2013], a survey taken by each participant by the end of the experiment. The results allow to understand whether the games made an impact on emotional state of the subjects, according to themselves. The results are represented by 7-factor structure. Five of them were further analysed, as they were the most relevant to the assumed game differences:

- Challenge – I felt time pressure/I had to put a lot effort,
- Tension – I was irritated/I feel angry,
- Negative affect – I felt bad/made me bored,

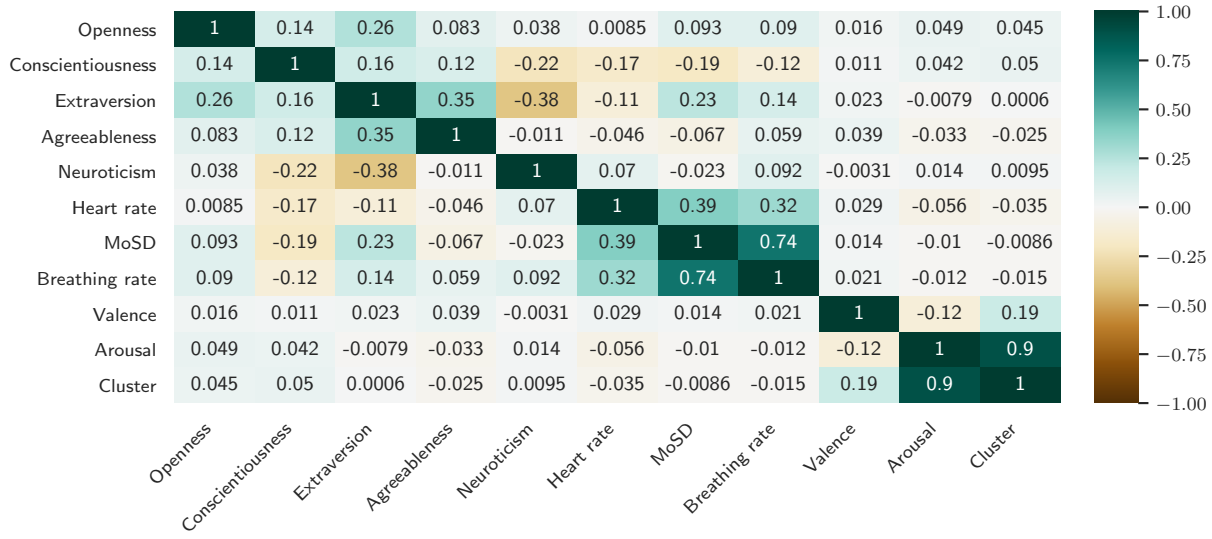


Figure 2: Correlation matrix for five personality traits, three ECG-related characteristics and widget responses represented as Valence, Arousal and Cluster.

Independent var.	df	MS	F	p
Conscientiousness	1	60.43	14.50	**< 0.001**
Openness	1	101.03	24.25	**< 0.001**
Agreeableness	1	44.51	10.68	**0.001**
Extraversion	1	9.61	2.31	0.13
Neuroticism	1	13.14	3.15	0.08
Heart rate	1	138.06	33.13	**< 0.001**
MoSD	1	11.27	2.71	0.10
Breathing rate	1	1.64	0.39	0.53
Residual Error	10881	4.17		

Table 4: ANOVA model for *arousal* with five personality traits and three ECG-related characteristics as independent variables.

- Positive affect – I felt good/made me happy,
- Competence – I felt competent/skillful.

The factors were compared to each other in order to dig into the feelings of players. The expectations for the first game were that subject is supposed to feel happy (high positive, low negative, low tension) and not challenged (high competence, low challenge). The second stage's purpose was contrary to the first one – high negative, tension and challenge, with low competence and positive. The huge difference is more likely to have an impact, as the contrast is hitting the player suddenly. Based on the GEQ results (see Fig. 4), one can state that everything worked as planned.

The Competence line during first gameplay was set pretty high, while leaving the tension line in the bottom, making the subject feel calm enough to let their guard down, but still be entertained by the gameplay. The second stage's extreme difficulty and pressure-building environment made the experience hard to enjoy. A very similar result can be seen in Negative/Tension comparison. About 95% of the participants agreed that the second level has left them irritated, 83,5%

Figure 3: Map for Stage 1 recreated from game logs with all death-related events marked as dots.

were not happy during and after the game. This cannot be said about the first stage, where according to the answers, only 30% of subjects felt somewhat irritated. Same outcome can be said about positive feedback for both stages – the first was keeping the emotions of participants on a very high level of happiness, while the second one changed it for a little one.

4 Discussion and Lessons Learned

As a summary of the analyses presented, we propose a set of guidelines concerning the issues one should pay attention to when creating games with the intention of using them as context-rich experimental environments:

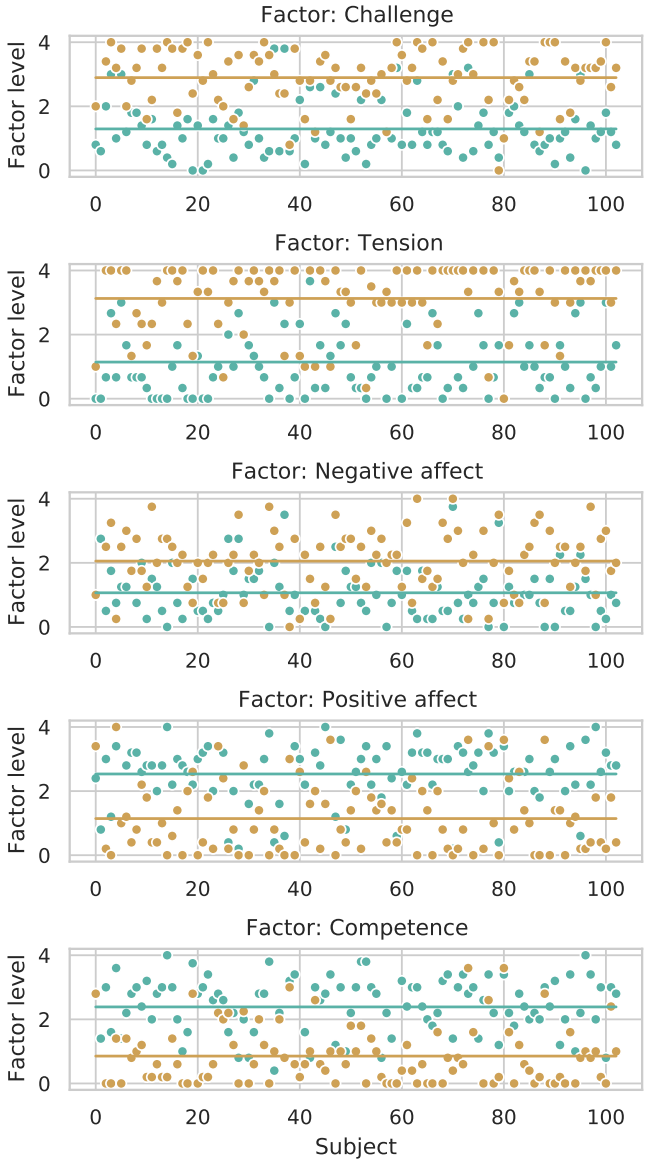


Figure 4: GEQ factors for the first and second level (green and yellow, respectively). Horizontal lines mark the average values.

1. *It is important to take into account the features of the subjects in the contextual information set.* In line with the results obtained from the BIRAFFE1 [Kutt *et al.*, 2021] and DEAP [Zhao *et al.*, 2019] datasets, the analyses summarised in Sect. 2 indicate interesting relationships between personality traits and physiological signals. Merging such several subject-related contextual information will allow a more accurate analysis leading to better modelling of a person's behaviour in the considered environment.

2. *The set of stimuli should be well balanced* so that there are neither too many (which will make analysis difficult) nor too few (the environment will not be interesting for the subject). Small levels, each focusing on selected aspects, should be preferred to one large level that combines all experimental manipulations. The levels analysed achieved their objectives well, as shown by the results of the GEQ questionnaire in Sect. 3.

3. *Logs should be collected as densely as possible, according to the specifics of the game being developed.* All features necessary to reproduce the gameplay should be recorded. In the analyses carried out, it was found that the logs were sufficiently detailed to reproduce the progress of the game. However, the data lacked information on the type of death in the second level, which would be useful to compare with the emotions felt at the time of death. This information is still reproducible, e.g., from the recorded screencasts, however it will require a fair amount of data processing.

4. *Maps with events marked on them are a useful tool for exploratory analysis of game logs.* There are a number of studies concerning the analysis of game logs (e.g., [Cheong *et al.*, 2008]), including those related to the evaluation of social science theories [Shim *et al.*, 2011]. However, to the best of our knowledge, data visualisation in the form of maps (as in Fig. 3) has not been done as part of the analyses. We believe that this is a valuable approach to quickly assess the validity of the data and to propose hypotheses that have not been considered before.

These findings will be incorporated into the preparation of the next experiment in the BIRAFFE series, planned for Autumn 2021.

Acknowledgements

The research has been supported by a grant from the Priority Research Area Digiworld under the Strategic Programme Excellence Initiative at the Jagiellonian University.

The authors are also grateful to Academic Computer Centre CYFRONET AGH and Jagiellonian University for granting access to the computing infrastructure built in the projects No. POIG.02.03.00-00-028/08 "PLATON – Science Services Platform" and No. POIG.02.03.00-00-110/13 "Deploying high-availability, critical services in Metropolitan Area Networks (MAN-HA)".

References

[Cheong *et al.*, 2008] Yun-Gyung Cheong, Arnav Jhala, Byung-Chull Bae, and Robert Michael Young. Automatically generating summary visualizations from game logs. In Christian Darken and Michael Mateas, editors, *AIIDE 2008*. The AAAI Press, 2008.

[Dzedzickis *et al.*, 2020] Andrius Dzedzickis, Arturas Kaklauskas, and Vytautas Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020.

[IJsselsteijn *et al.*, 2013] Wijnand A. IJsselsteijn, Yvonne A. W. de Kort, and Karolien Poels. *The Game Experience Questionnaire*. Technische Universiteit Eindhoven, 2013.

[Kutt *et al.*, 2020] Krzysztof Kutt, Dominika Drążyk, Maciej Szelążek, Szymon Bobek, and Grzegorz J. Nalepa. The BIRAFFE2 experiment. study in bio-reactions and faces for emotion-based personalization for AI systems. *CoRR*, abs/2007.15048, 2020.

[Kutt *et al.*, 2021] Krzysztof Kutt, Dominika Drążyk, Szymon Bobek, and Grzegorz J. Nalepa. Personality-based affective adaptation methods for intelligent systems. *Sensors*, 21(1):163, 2021.

[Nalepa *et al.*, 2019] Grzegorz J. Nalepa, Krzysztof Kutt, and Szymon Bobek. Mobile platform for affective context-aware systems. *Future Generation Computer Systems*, 92:490–503, mar 2019.

[Prinz, 2006] Jesse J. Prinz. *Gut Reactions. A Perceptual Theory of Emotion*. Oxford University Press, Oxford, 2006.

[Shim *et al.*, 2011] Kyong Jin Shim, Nishith Pathak, Muhammad Aurangzeb Ahmad, Colin DeLong, Zoheb Borbora, Amogh Mahapatra, and Jaideep Srivastava. Analyzing human behavior from multiplayer online game logs: A knowledge discovery approach. *IEEE Intell. Syst.*, 26(1):85–89, 2011.

[van Gent *et al.*, 2019] Paul van Gent, Haneen Farah, Nicole van Nes, and Bart van Arem. Analysing noisy driver physiology real-time using off-the-shelf sensors: Heart rate analysis software from the taking the fast lane project. *Journal of Open Research Software*, 7(1):32, 2019.

[Zhao *et al.*, 2019] Sicheng Zhao, Amir Gholaminejad, Guiguang Ding, Yue Gao, Jungong Han, and Kurt Keutzer. Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM Trans. Multim. Comput. Commun. Appl.*, 15(1s):14:1–14:18, 2019.

[Żuchowska *et al.*, 2020] Laura Żuchowska, Krzysztof Kutt, Krzysztof Geleta, Szymon Bobek, and Grzegorz J. Nalepa. Affective games provide controlable context. proposal of an experimental framework. In Jörg Cassens, Rebekah Wegener, and Anders Kofod-Petersen, editors, *Proceedings of the Eleventh International Workshop Modelling and Reasoning in Context co-located with the 24th European Conference on Artificial Intelligence, MRC@ECAI 2020, Santiago de Compostela, Galicia, Spain, August 29, 2020*, volume 2787 of *CEUR Workshop Proceedings*, pages 45–50. CEUR-WS.org, 2020.

Bi-ISCA: Bidirectional Inter-Sentence Contextual Attention Mechanism for Detecting Sarcasm in User Generated Noisy Short Text

Prakamya Mishra[1*], **Saroj Kaushik**[2], **Kuntal Dey**[3]
[1],[2]Shiv Nadar University
[3]Accenture Technology Labs
{pm669, saroj.kaushik}@snu.edu.in, kuntal.dey@accenture.com

Abstract

Many online comments on social media platforms are hateful, humorous, or sarcastic. The sarcastic nature of these comments (especially the short ones) alters their actual implied sentiments, which leads to misinterpretations by the existing sentiment analysis models. A lot of research has already been done to detect sarcasm in the text using user-based, topical, and conversational information but not much work has been done to use inter-sentence contextual information for detecting the same. This paper proposes a new deep learning architecture that uses a novel Bidirectional Inter-Sentence Contextual Attention mechanism (Bi-ISCA) to capture inter-sentence dependencies for detecting sarcasm in the user-generated short text using only the conversational context. The proposed deep learning model demonstrates the capability to capture explicit, implicit, and contextual incongruous words & phrases responsible for invoking sarcasm. Bi-ISCA generates results comparable to the state-of-the-art on two widely used benchmark datasets for the sarcasm detection task (Reddit and Twitter). To the best of our knowledge, none of the existing models use an inter-sentence contextual attention mechanism to detect sarcasm in the user-generated short text using only conversational context.

1 Introduction

Sentiment analysis is one of the most important natural language processing (NLP) applications. Its goal is to identify, extract, quantify, and study subjective information. The sudden rise in the usage of social media platforms as a means of communication has led to a vast amount of data being shared between its users on a wide range of topics. This type of data is very helpful to several organizations for analyzing the sentiments of people towards products, movies, political events, etc. Understanding the unique intricacies of the human language remains one of the most important pending NLP problems of this time. Humans regularly use sarcasm as a crucial part of the day-to-day conversations when venting, arguing, or maybe engaging on social media platforms. Sarcastic remarks on these platforms inflict problems on the existing sentiment analysis systems in identifying the true intentions of the users.

The Cambridge Dictionary[1] describes sarcasm as an irony conveyed hilariously or amusingly to criticize something. Sarcasm may not show criticism on the surface but instead might have a criticizing implied meaning. Such a figurative aspect of sarcasm makes it difficult to be detected in the modern micro texts [Ghosh and Veale, 2016]. Several linguistic research has been done to analyze different aspects of sarcasm. Kind of responses evoked because of comments has been considered a major indicator of sarcasm [Eisterhold *et al.*, 2006]. [Wilson, 2006] states that circumstantial incongruity between a comment and its corresponding contextual information plays an important role in implying sarcasm.

Previous research works have used policy-based, statistical, and deep-learning-based methods for detecting sarcasm. The use of contextual information like conversational context, author personality features, or prior knowledge of the topic, have proved to be very useful. [Khattri *et al.*, 2015] used sentiments of the author's historical tweets as context. [Rajadesingan *et al.*, 2015] used personality features like the author's familiarity with twitter, language (structure and word usage), and the author's familiarity with sarcasm (history of previous sarcastic tweets) for consolidating context. [Bamman and Smith, 2015] explored the use of historical terms, topics, and sentiments along with profile information as the author's context. They also exploited the use of conversational context like the immediate previous tweets in the thread. [Joshi *et al.*, 2015] demonstrated that concatenation of preceding comment with the objective comment in a discussion forum led to an increase in the precision score.

Overall in recent years a lot of work has been done to use different types of contextual information for sarcasm detection but none of them have used inter-sentence dependencies. In this paper, we propose a novel Bidirectional Inter-Sentence Contextual Attention mechanism (Bi-ISCA) based deep learning neural network for sarcasm detection. The main contribution of this paper can be summarised as follows:

- We propose a new deep learning architecture that uses a novel Bidirectional Inter-Sentence Contextual attention mechanism (Bi-ISCA) for detecting sarcasm in short texts

*Contact Author

[1]https://dictionary.cambridge.org/

(short texts are more difficult to analyze due to shortage of contextual information).

- Bi-ISCA focuses on only using the conversational contextual comment/tweet for detecting sarcasm rather than using any other topical/personality-based features, as using only the contextual information enriches the model's ability to capture syntactical and semantical textual properties responsible for invoking sarcasm.
- We also explain model behavior and predictions by visualizing attention maps generated by Bi-ISCA, which helps in identifying significant parts of the sentences responsible for invoking sarcasm.

The rest of the paper is organized as follows. Section 2 describes the related work. Then section 3, explains the proposed model architecture for detecting sarcasm. Section 4 will describe the datasets used, pre-processing pipeline, and training details for reproducibility. Then experimental results are explained in section 5 and section 6 illustrates model behavior and predictions by visualizing attention maps. Finally we conclude in section 7.

2 Related Work

A diverse spectrum of approaches has been used to detect sarcasm. Recent sarcasm detection approaches have either mainly focused on using machine learning based approaches that leverage the use of explicitly declared relevant features or they focus on using neural network based deep learning approaches that do not require handcrafted features. Also, the recent advances in using deep learning for preforming natural language processing tasks have led to a promising increase in the performance of these sarcasm detection systems.

A lot of research has been done using bag of words as features. However, to improve performance, scholars started to explore the use of several other semantic and syntactical features like punctuations [Tsur *et al.*, 2010]; emotion marks and intensifiers [Liebrecht *et al.*, 2013]; positive verbs and negative phrases [Riloff *et al.*, 2013]; polarity skip grams [Reyes *et al.*, 2013]; synonyms & ambiguity[Barbieri *et al.*, 2014]; implicit and explicit incongruity-based [Joshi *et al.*, 2015]; sentiment flips [Rajadesingan *et al.*, 2015]; affect-based features derived from multiple emotion lexicons [Farías *et al.*, 2016].

Every day an enormous amount of short text data is generated by users on popular social media platforms like Twitter[2] and Reddit[3]. Easy accessibility of such data sources has enticed researchers to use them for extracting user-based and discourse-based features. [Hazarika *et al.*, 2018] utilized contextual information by making user-embeddings for capturing indicative behavioral traits. These user-embeddings incorporated personality features along with the author's writing style (using historical posts). They also used discourse comments along with background cues and topical information for detecting sarcasm. They performed their experiments on the largest Reddit dataset SARC [Khodak *et al.*, 2018]. Many have only used the target text for classification purposes, where a target text is a textual unit that has to be classified as sarcastic or not. Simply using gated recurrent units (GRU) [Cho *et al.*, 2014] or long short term memory (LSTM) [Hochreiter and Schmidhuber, 1997] do not capture in between interactions of word pairs which makes it difficult to model contrast and incongruity. [Tay *et al.*, 2018] were able to solve this problem by looking in-between word pairs using a multi-dimensional intra-attention recurrent network. They focused on modeling the intra-sentence relationships among the words. [Kumar *et al.*, 2020] exploited the use of a multi-head attention mechanism [Vaswani *et al.*, 2017] which could capture dependencies between different representations subspaces in different positions. Their model consisted of a word encoder for generating new word representations by summarizing comment contextual information in a bidirectional manner. On top of that, they used multi-head attention for focusing on different contexts of a sentence, and in the end, a simple multi-layer perceptron was used for classification.

There has not been much work done in conversation dependent (comment and reply) approaches for sarcasm detection. [Ghaeini *et al.*, 2018] proposed a model that not only used information from the target utterance but also used its conversational context to perceive sarcasm. They aimed to detect sarcasm by just using the sequences of sentences, without any extra knowledge about the user and topic. They combined the predictions from utterance-only and conversation-dependent parts for generating its final prediction which was able to capture the words responsible for delivering sarcasm. [Ghosh and Veale, 2017] also modeled conversational context for sarcasm detection. They also attempted to derive what parts of the conversational context triggered a sarcastic reply. Their proposed model used sentence embeddings created by taking an average of word embeddings and a sentence-level attention mechanism was used to generate attention induced representations of both the context and the response which was later concatenated and used for classification.

Among all the previous works, [Ghaeini *et al.*, 2018] and [Ghosh and Veale, 2017] share similar motives of detecting sarcasm using only the conversational context. However, we introduce a novel Bidirectional Inter-Sentence Contextual Attention mechanism (Bi-ISCA) for detecting sarcasm. Unlike previous works, our work considers short texts for detecting sarcasm, which is far more challenging to detect when compared to long texts as long texts provide much more contextual information.

3 Model

This section will introduce the proposed Bi-ISCA: Bidirectional Inter Sentence Contextual Attention based neural network for sarcasm detection (as shown in Figure 1). Sarcasm detection is a binary classification task that tries to predict whether a given comment is sarcastic or not. The proposed model uses comment-reply pairs for detecting sarcasm. The input to the model is represented by $U = [W_1^u, W_2^u, ..., W_n^u]$ and $V = [W_1^v, W_2^v, ..., W_n^v]$, where U represents the comment sentence and V represents the reply sentence (both sentences padded to a length of n). Here, $W_i^u, W_j^v \in \mathbb{R}^d$ are $d-$dimensional word embedding vectors. The objective is

[2]www.twitter.com/

[3]www.reddit.com/

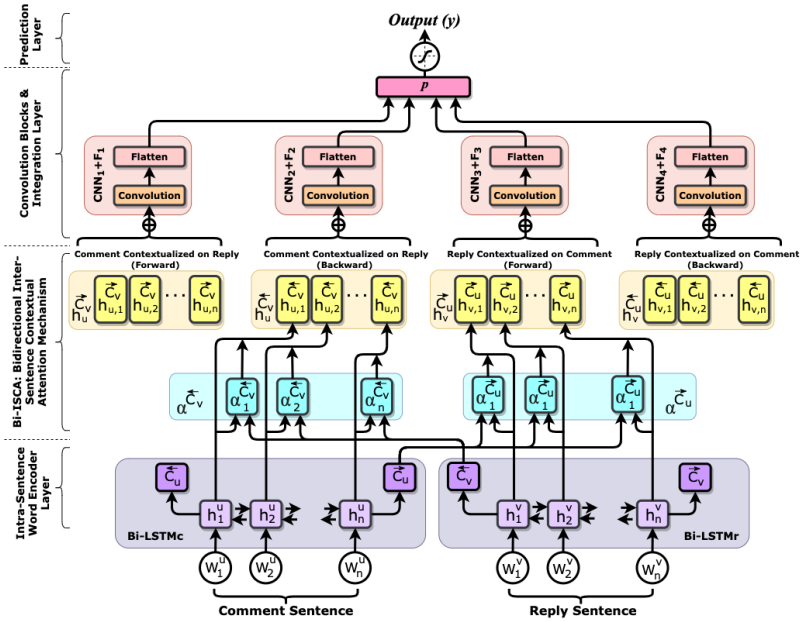


Figure 1: Bi-ISCA: Bi-Directional Inter-Sentence Contextual Attention Mechanism for Sarcasm Detection.

to predict label y which indicates whether the reply to the corresponding comment was sarcastic or not.

3.1 Intra-Sentence Word Encoder Layer

The primary purpose of this layer is to summarize intra-sentence contextual information from both directions in both the sentences (comment & reply) using Bidirectional Long Short Term Memory Networks (Bi-LSTM). A Bi-LSTM [Schuster and Paliwal, 1997] processes information in both the directions using a forward LSTM [Hochreiter and Schmidhuber, 1997] $\overrightarrow{h}$, that reads the sentence $S = [w_1, w_2, ..., w_n]$ from w_1 to w_n and a backward LSTM $\overleftarrow{h}$ that reads the sentence from w_n to w_1. Hidden states from both the LSTMs are added to get the final hidden state representations of each word. So the hidden state representation of the t^{th} word (h_t) can be represented by the sum of t^{th} hidden representations of the forward and backward LSTMs ($\overrightarrow{h_t}$,$\overleftarrow{h_t}$) as show in equations below.

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(w_t, \overrightarrow{h_{t-1}}); \overleftarrow{h_t} = \overleftarrow{LSTM}(w_t, \overleftarrow{h_{t-1}}) \quad (1)$$

$$h_t = \overleftarrow{h_t} + \overrightarrow{h_t} \quad (2)$$

This Intra-Sentence Word Encoder Layer consists of two independent Bidirectional LSTMs for both comment ($BiLSTM_c$) and reply ($BiLSTM_r$). Apart from the hidden states, both these Bi-LSTMs also generate separate (forward & backward) final cell states represented by $\overleftarrow{C}$ & $\overrightarrow{C}$. The comment sentence U is given as an input to $BiLSTM_c$ and the reply sentence V is given as an input to $BiLSTM_r$. The outputs of both the Bi-LSTMs are represented by the equations 3 and 4.

$$\overrightarrow{C_u}, h^u, \overleftarrow{C_u} = BiLSTM_c(U) \quad (3)$$

$$\overrightarrow{C_v}, h^v, \overleftarrow{C_v} = BiLSTM_r(V) \quad (4)$$

Here, $\overrightarrow{C_u}, \overrightarrow{C_v} \in \mathbb{R}^d$ are the final cell states of the forward LSTMs corresponding to $BiLSTM_c$ & $BiLSTM_r$; $\overleftarrow{C_u}, \overleftarrow{C_v} \in \mathbb{R}^d$ are the final cell states of the backward LSTMs corresponding to $BiLSTM_c$ & $BiLSTM_r$; $h^u = [h^u_1, h^u_2, ..., h^u_n]$ and $h^v = [h^v_1, h^v_2, ..., h^v_n]$ are the hidden state representations of $BiLSTM_c$ & $BiLSTM_r$ respectively, where $h^u_i, h^v_j \in \mathbb{R}^d$ and $h^u, h^v \in \mathbb{R}^{n \times d}$.

3.2 Bi-ISCA: Bidirectional Inter-Sentence Contextual Attention Mechanism

Sarcasm is context-dependent in nature. Even humans sometimes have a hard time understanding sarcasm without having any contextual information. The hidden states generated by both the Bi-LSTMs ($BiLSTM_c$ & $BiLSTM_r$) captures the intra-sentence bidirectional contextual information in comment & reply respectively, but fails to capture the inter-sentence contextual information between them. This paper introduces a novel Bidirectional Inter-Sentence Contextual Attention mechanism (Bi-ISCA) for capturing the inter-sentence contextual information between both the sentences.

Bi-ISCA uses hidden state representations of U & V along with the auxiliary sentence's cell state representations ($\overrightarrow{C}$ & $\overleftarrow{C}$) to capture the inter-sentence contextual information. At first, the attention mechanism captures four sets of attentions scores namely, ($\alpha^{\overrightarrow{C_u}}, \alpha^{\overleftarrow{C_u}}, \alpha^{\overrightarrow{C_v}}, \alpha^{\overleftarrow{C_v}} \in \mathbb{R}^n$). These sets of inter-sentence attention scores are used to generate new inter-sentence contextualized hidden representations. Then ($\alpha^{\overrightarrow{C_u}}, \alpha^{\overleftarrow{C_u}}$) are calculated using the hidden state representations of $BiLSTM_r$ along with the forward and backward final states ($\overrightarrow{C_u}, \overleftarrow{C_u}$) of $BiLSTM_c$ (as shown in equations

5 & 6), similarly ($\alpha^{\overrightarrow{C_v}}, \alpha^{\overleftarrow{C_v}}$) are calculated using the hidden state representations of $BiLSTM_c$ along with the forward and backward final states ($\overrightarrow{C_v}, \overleftarrow{C_v}$) of $BiLSTM_r$ (as shown in equations 7 & 8). In the equations below ($\bullet$) represents a dot product between two vectors.

$$\alpha^{\overrightarrow{C_u}} = [\alpha_1^{\overrightarrow{C_u}}, \alpha_2^{\overrightarrow{C_u}}, ..., \alpha_n^{\overrightarrow{C_u}}]; \alpha_i^{\overrightarrow{C_u}} = \overrightarrow{C_u} \bullet h_i^v \quad (5)$$

$$\alpha^{\overleftarrow{C_u}} = [\alpha_1^{\overleftarrow{C_u}}, \alpha_2^{\overleftarrow{C_u}}, ..., \alpha_n^{\overleftarrow{C_u}}]; \alpha_i^{\overleftarrow{C_u}} = \overleftarrow{C_u} \bullet h_i^v \quad (6)$$

$$\alpha^{\overrightarrow{C_v}} = [\alpha_1^{\overrightarrow{C_v}}, \alpha_2^{\overrightarrow{C_v}}, ..., \alpha_n^{\overrightarrow{C_v}}]; \alpha_i^{\overrightarrow{C_v}} = \overrightarrow{C_v} \bullet h_i^u \quad (7)$$

$$\alpha^{\overleftarrow{C_v}} = [\alpha_1^{\overleftarrow{C_v}}, \alpha_2^{\overleftarrow{C_v}}, ..., \alpha_n^{\overleftarrow{C_v}}]; \alpha_i^{\overleftarrow{C_v}} = \overleftarrow{C_v} \bullet h_i^u \quad (8)$$

In the next step, the above calculated sets of inter-sentence attention scores $\alpha^{\overrightarrow{C_u}}, \alpha^{\overleftarrow{C_u}}$) are multiplied back with the hidden state representations of $BiLSTM_r$ to generate two new set of hidden representations $h_v^{\overrightarrow{C_u}}, h_v^{\overleftarrow{C_u}} \in \mathbb{R}^{n \times d}$ of the reply sentence namely, reply contextualized on comment (forward) & reply contextualized on comment (backward) respectively (as shown in equations 9 & 10). Similarly $\alpha^{\overrightarrow{C_v}}, \alpha^{\overleftarrow{C_v}}$ are multiplied back with the hidden state representations of $BiLSTM_c$ to generate two new set of hidden representations $h_u^{\overrightarrow{C_v}}, h_u^{\overleftarrow{C_v}} \in \mathbb{R}^{n \times d}$ of the comment sentence namely, comment contextualized on reply (forward) & comment contextualized on reply (backward) respectively (as shown in equations 11 & 12). In the equations below ($\times$) represents multiplication between a scalar and a vector.

$$h_v^{\overrightarrow{C_u}} = [h_{v,1}^{\overrightarrow{C_u}}, h_{v,2}^{\overrightarrow{C_u}}, ..., h_{v,n}^{\overrightarrow{C_u}}], ; h_{v,i}^{\overrightarrow{C_u}} = \alpha_i^{\overrightarrow{C_u}} \times h_i^v \quad (9)$$

$$h_v^{\overleftarrow{C_u}} = [h_{v,1}^{\overleftarrow{C_u}}, h_{v,2}^{\overleftarrow{C_u}}, ..., h_{v,n}^{\overleftarrow{C_u}}], ; h_{v,i}^{\overleftarrow{C_u}} = \alpha_i^{\overleftarrow{C_u}} \times h_i^v \quad (10)$$

$$h_u^{\overrightarrow{C_v}} = [h_{u,1}^{\overrightarrow{C_v}}, h_{u,2}^{\overrightarrow{C_v}}, ..., h_{u,n}^{\overrightarrow{C_v}}], ; h_{u,i}^{\overrightarrow{C_v}} = \alpha_i^{\overrightarrow{C_v}} \times h_i^u \quad (11)$$

$$h_u^{\overleftarrow{C_v}} = [h_{u,1}^{\overleftarrow{C_v}}, h_{u,2}^{\overleftarrow{C_v}}, ..., h_{u,n}^{\overleftarrow{C_v}}], ; h_{u,i}^{\overleftarrow{C_v}} = \alpha_i^{\overleftarrow{C_v}} \times h_i^u \quad (12)$$

3.3 Integration and Final Prediction

The proposed model uses Convolutional Neural Networks (CNN) [Lecun *et al.*, 1998] for capturing location-invariant local features from the newly obtained contextualized hidden representations $h_u^{\overleftarrow{C_v}}, h_u^{\overrightarrow{C_v}}, h_v^{\overleftarrow{C_u}}, h_v^{\overrightarrow{C_u}}$. Four independent CNN blocks ($CNN_1, CNN_2, CNN_3, CNN_4$) are used, corresponding to each of the newly obtained contextualized hidden representations. Each CNN block consists two convolutional layers. Both the convolution layer consist of k filters of height h. The role of these filters is to detect particular features at different locations of the input. The output c_i^l of the l^{th} layer consists of k^l feature maps of height h. The i^{th} feature map (c_i^l) is calculated as:

$$c_i^l = b_i^l + \sum_{k^{l-1}}^{j=1} K_{i,j}^l * c_j^{l-1} \quad (13)$$

In the above equation, b_i^l is a bias matrix and $K_{i,j}^l$ is a filter connecting j^{th} feature map of layer $(l-1)$ to the i^{th} feature map of layer (l). The output of each convolution layer is passed through a activation function f. The proposed model uses *Leaky*ReLu as its activation function.

$$f = \begin{cases} a * x, & \text{for } x \geq 0; a \in \mathbb{R} \quad (14) \\ x, & \text{for } x < 0 \quad (15) \end{cases}$$

For each of the CNN blocks, the corresponding contextualized hidden representations are first concatenated ($\oplus$) and then given as input. The outputs of all the CNN blocks are flattened ($F_1, F_2, F_3, F_4 \in \mathbb{R}^{dk}$) and concatenated to generate a new vector ($p \in \mathbb{R}^{4dk}$), where d represents the dimension of the hidden representation and k represents number of convolutional filters used. This concatenated (p) vector is then given as input to a dense layer having $4dk$ neurons and is followed by the final sigmoid prediction layer.

$$F_1 = CNN_1([h_{u,1}^{\overrightarrow{C_v}} \oplus h_{u,2}^{\overrightarrow{C_v}} \oplus .... \oplus h_{u,n}^{\overrightarrow{C_v}}]) \quad (16)$$

$$F_2 = CNN_2([h_{u,1}^{\overleftarrow{C_v}} \oplus h_{u,2}^{\overleftarrow{C_v}} \oplus .... \oplus h_{u,n}^{\overleftarrow{C_v}}]) \quad (17)$$

$$F_3 = CNN_3([h_{v,1}^{\overrightarrow{C_u}} \oplus h_{v,2}^{\overrightarrow{C_u}} \oplus .... \oplus h_{v,n}^{\overrightarrow{C_u}}]) \quad (18)$$

$$F_4 = CNN_4([h_{v,1}^{\overleftarrow{C_u}} \oplus h_{v,2}^{\overleftarrow{C_u}} \oplus .... \oplus h_{v,n}^{\overleftarrow{C_u}}]) \quad (19)$$

$$p = [F_1 \oplus F_2 \oplus F_3 \oplus F_4] \quad (20)$$

$$\hat{y} = \sigma(Wp + b), \quad W \in \mathbb{R}^{4dk}; b \in \mathbb{R} \quad (21)$$

The proposed model uses the binary cross-entropy as the training loss function as shown in equation 22. Here (L) is the cost function, $\hat{y}_i \in \mathbb{R}$ represents the output of the proposed model, $y_i \in \mathbb{R}$ represents the true label and $N \in \mathbb{N}$ represents the number of training samples.

$$L = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i) \quad (22)$$

4 Evaluation Setup

4.1 Dataset

This paper focuses on detecting sarcasm in the user-generated short text using only the conversational context. Social media platforms like Reddit and Twitter are widely used by users for posting opinions and replying to other's opinions. They have proved to be of a great source for extracting conversational data. So the experiments were conducted on two publicly available benchmark datasets (Reddit & Twitter) used for the sarcasm detection task. Both the datasets consist of comments and reply pairs.

SARC[4] **Reddit [Khodak *et al.*, 2018]** is the largest dataset available for sarcasm detection containing millions of sarcastic/non-sarcastic comments-reply pairs from the social media site Reddit. This dataset was generated by scraping

[4]https://nlp.cs.princeton.edu/SARC/2.0/

			No. of comment-reply pairs		Avg. no. of words per comment		Avg. no. of words per reply	
			Sarcastic	Non-Sarcastic	Sarcastic	Non-Sarcastic	Sarcastic	Non-Sarcastic
Training set	Reddit	Balanced	81205	81205	12.69	12.67	12.19	12.21
		Imbalanced	16303	81205	12.69	12.65	12.15	12.21
	Twitter	Balanced	3496	3496	24.97	24.97	24.25	24.25
Testing set	Reddit	Balanced	9058	9058	12.71	12.64	12.14	12.22
		Imbalanced	1747	9058	12.73	12.69	12.20	12.21
	Twitter	Balanced	874	874	24.97	24.97	24.25	24.25

Table 1: Statics of the SARC dataset and FigLang 2020 workshop Twitter dataset.

comments from Reddit containing the \s (sarcasm) tag. It contains replies, their parent comment (acts as context), and a label that shows whether the reply was sarcastic/non-sarcastic to their corresponding parent comment. To compare the performance of the model on a different dataset (latest), the proposed model was also evaluated on the Twitter dataset provided in the **FigLang**[5] **2020 workshop [Ghosh *et al.*, 2020]** for the "sarcasm detection shared task". This consists of sarcastic/non-sarcastic tweets and their corresponding contextual parent tweets. The sarcastic tweets were collected using hashtags like #sarcasm, #sarcastic, and #irony, similarly non-sarcastic tweets were collected using hashtags like #happy, #sad, and #hate. This dataset sometime contains more than one contextual parent tweet, so in those cases, all of the contextual tweets are considered independently with the target tweet.

In both the datasets, replies are the target comment/tweet to be classified as sarcastic/non-sarcastic, and their corresponding parent comment/tweet acts as context. Both the datasets constitute of comments/tweets of varying lengths, but because this paper only focuses on detecting sarcasm in the short text, only the short comment/reply pairs were used. Comment/reply sentences of length (no. of words) less than 20, 40 were used in the case of SARC and Twitter dataset respectively. In both cases, the balanced datasets contain equal proportions of sarcastic/non-sarcastic comment/reply pairs, and the imbalanced datasets maintain a 20:80 ratio (approximately) between sarcastic and non-sarcastic comment/reply pairs. Testing was done on 10% of the dataset and the rest was used for training. 10% of the training set was used for validation purposes. Statistics of both the datasets are shown in Table 1.

4.2 Data Preprocessing

The preprocessing of the textual data was done by first lowercasing all the sentences and separating punctuations from the words. We do not remove the stop-words because we believe that sometimes stop-words play a major role in making a sentence sarcastic e.g., *"is it?"* and *"am I?"*. The problem with social media platforms is that, users use a lot of abbreviations, shortened words and slang words like, *"IMO"* for *"in my opinion"*, *lmk"* for *"let me know "*, *"fr"* for *"for"*, etc. These words are challenging to taken care of in the NLP tasks, particularly in the automatic discovery of flexible word usages. So to solve this problem, these words are converted to their corresponding full-forms using abbreviation/slang word dictionaries obtained from urban dictionary[6]. After this, all the sentences were tokenized into a list of words. The proposed model had a fixed input size for both comment and reply, but not all the sentences were of the same length. So all the sentences were padded to the length of the longest sentence (20 in the case of the Reddit dataset and 40 in the case of the Twitter dataset). Word embeddings are used to give semantically-meaningful dense representations to the words. Word-based embeddings are constructed using contextual words whereas character-based embeddings are constructed from character n-grams of the words. Character-based in contrast to the Word-based embeddings solves the problem of out of vocabulary words and performs better in the case of infrequent words by creating word embeddings based only on their spellings. So for generating proper representations for words we have used FastText[7], a character-based word embedding. This would not only give words better representation compared to the word-based model but also incorporate slang/shortened/infrequent words (which commonly appear in social media platforms).

4.3 Training Details

We have used macro-averaged (*F1*) and accuracy (*Acc*) scores as the evaluation metric, as it is standard for the sarcasm detection task. We have also reported Precision (*P*) and Recall (*R*) scores in the case of the Twitter dataset as well as for the Reddit dataset (wherever available). Hyperparameter tuning was used to find optimum values of the hyperparameters. The FastText embeddings used were of size $d = 30$ and were trained for 30 iterations having window size of 3, 5 in the case of SARC, and Twitter dataset respectively. The number of filters in all the convolutional blocks were [64, 64] of height [2, 2]. The learning optimizer used is Adam with an initial learning rate of 0.01. The value of α in all the *Leaky*ReLu layers was set to 0.3. All the models were trained for 20 epochs. L2 regularization set to 10^{-2} is applied to all the feed-forward connections along with early stopping having the patience of 5 to avoid overfitting. The mini-batch size was tuned amongst {100, 500, 1000, 2000, 3000, 4000} and was observed that mini-batch size of 2000, 500 gave the best performance for the SARC and Twitter dataset respectively.

The recent success of transformer-based language models has led to their wide usage in sentiment analysis tasks. They are known for generating high quality high dimensional word representations (768-dimensional for BERT). Their only drawback is that they require high processing power and memory to train. The above-mentioned configuration of the proposed model generates ≈1120K trainable parameters, and increasing either the embedding size or the number of tokens in a sentence led to an exponential increase in the number of trainable parameters. So due to computational resource limitations, we limited our experiments to lower-dimensional word embeddings.

[5] sites.google.com/view/figlang2020

[6] https://www.urbandictionary.com/

[7] https://fasttext.cc/

5 Results

Models	Balance				Imbalanced			
	Acc	*F1*	*P*	*R*	*Acc*	*F1*	*P*	*R*
CNN-SVM [Poria *et al.*, 2016] †⋆	68.0	68.0	–	–	69.0	79.0	–	–
AMR [Ghaeini *et al.*, 2018] ‡	69.5	69.5	74.8	69.7	–	–	–	–
[Ghosh and Veale, 2017] ‡	–	67.8	68.2	67.9	–	–	–	–
CUE-CNN [Amir *et al.*, 2016] †⋆	70.0	69.0	–	–	73.0	81.0	–	–
MHA-BiLSTM [Kumar *et al.*, 2020] †	–	77.5	72.6	83.0	–	56.8	60.3	53.7
CASCADE [Hazarika *et al.*, 2018] ‡⋆	77.0	77.0	–	–	79.0	86.0	–	–
CASCADE (only discourse features) ‡	68.0	66.0	–	–	68.0	78.0	–	–
Bi-ISCA (this paper) ‡	**72.3**	**75.7**	**74.2**	**77.6**	**71.9**	**74.4**	**73.0**	**75.8**
Δ increase w.r.t CASCADE (only discourse features)	4.3 ↑	9.7 ↑	–	–	3.9 ↑	3.6 ↓	–	–

† Uses only target sentence, ‡ Uses context along with target sentence,
⋆ Uses personality-based features

Table 2: Results on the SARC dataset. Models haveing only ‡ uses only contextual text for detecting sarcasm.

Bi-ISCA focuses on only using the contextual comment/tweet for detecting sarcasm rather than using any other topical/personality-based features. Using only the contextual information enriches the model's ability to capture syntactical and semantical textual properties responsible for invoking sarcasm in any type of conversation. Table 2 reports performance results on the SARC datasets. For comparison purposes, F1-score (*F1*), Accuracy score (*Acc*), Precision (*P*) and Recall (*R*) were used.

When compared with the existing works, Bi-ISCA was able to outperform all the models (only ‡) that use only conversational context for sarcasm detection (Improvement of Δ 7.9% in F1 score when compared to [Ghosh and Veale, 2017]; Δ 6.2% in F1 score and Δ 2.8% in accuracy when compared to AMR [Ghaeini *et al.*, 2018]), and was even able to perform better than the models (†⋆) that use personality-based features along with the target sentence for detecting sarcasm (improvement of Δ 7.7% in F1 and Δ 4.3% in accuracy score when compared to CNN-SVM [Poria *et al.*, 2016]; Δ 6.7% in F1 score and Δ 2.3% in accuracy when compared to CUE-CNN [Amir *et al.*, 2016]). MHA-BiLSTM [Kumar *et al.*, 2020] had a Δ 1.8% higher F1 score in the balanced dataset but Bi-ISCA was able to show drastic improvement of Δ 17.6% in the imbalanced dataset, which demonstrated the ability of Bi-ISCA to handle class imbalance.

The current state-of-the-art on the SARC dataset is achieved by CASCADE. Even though CASCADE uses personality-based features and contextual information along with large sentences of average length ≈55-62 (very large compared to our dataset, which gives them the advantage of using a lot more contextual information), Bi-ISCA was able to achieve an F1 score comparable to it (despite using relatively short text). In comparison with CASCADE that only uses discourse-based features, Bi-ISCA performed drastically better with an increase of Δ 9.7% in F1 and Δ 4.3% in accuracy score for the balanced dataset.

Bi-ISCA clearly demonstrated its capabilities to robustly handle an imbalance in the dataset, although it was unable to outperform both the CASCADE models. This slightly poor performance in the imbalanced dataset can be explained by the length of sentences used by CASCADE, which are significantly (≈5 times) greater than the ones on which Bi-ISCA was tested. Longer sentences result in increased contextual information which improves performance especially in the case of imbalance where little extra information can lead to a drastic increase in performance.

Models	P	R	F1
Baseline ($LSTM_attn$)	70.0	66.9	68.0
BERT-Large+BiLSTM+SVM [Baruah *et al.*, 2020]	73.4	73.5	73.4
BERT+CNN+LSTM [Srivastava *et al.*, 2020]	74.2	74.6	74.1
RoBERTa+LSTM [Kumar and Anand, 2020]	77.3	77.4	77.2
RoBERT-Large [Dong *et al.*, 2020]	79.1	79.4	79.0
RoBERT+Multi-Initialization Ensemble [Jaiswal, 2020]	79.2	79.3	79.1
BERT + BiLSTM + NeXtVLAD + Context Ensemble + Data Augmentation [Lee *et al.*, 2020]	93.2	93.6	93.1
Bi-ISCA (this paper)	**89.4**	**94.8**	**91.7**

Table 3: Results on the FigLang 2020 workshop Twitter dataset.

Table 3 reports Precision (*P*), Recall (*R*), and F1-score (*F1*) of different models from the leaderboard of FigLang 2020 sarcasm detection shared task using the Twitter dataset. In this case, not only Bi-ISCA was able to outperform the baseline model [Ghosh *et al.*, 2020] (improvement of Δ 19.4%, Δ 27.9% & Δ 23.7% in precision, recall, and F1 score respectively), but was also able to perform comparably to the state-of-the-art [Lee *et al.*, 2020] with a Δ 1.2% increase in recall, which further validates the performance of the proposed model. Even though all the models other than the baseline in Table 3 are a transformer-based model, Bi-ISCA was able to outperform them all.

6 Discussion

1.	*CcR* ***(R)***:	dog traumatized by abuse is caressed for the first tim
	CcR ***(L)***:	dog traumatized by abuse is caressed for the first tim
	RcC ***(R)***:	what amazing idea to force a dog to be pet while cornered when its terrified
	RcC ***(L)***:	what amazing idea to force a dog to be pet while cornered when its terrified
2.	*CcR* ***(R)***:	no pay gap between men and women at microsoft company say
	CcR ***(L)***:	no pay gap between men and women at microsoft company say
	RcC ***(R)***:	m also announces that windows never ever crashes and has security holes
	RcC ***(L)***:	m also announces that windows never ever crashes and has security holes
3.	*CcR* ***(R)***:	its totally okay to treat autonomous ton vehicles with caution
	CcR ***(L)***:	its totally okay to treat autonomous ton vehicles with caution
	RcC ***(R)***:	when im blind drunk and get behind the wheel its as though its an autonomous vehicle
	RcC ***(L)***:	when im blind drunk and get behind the wheel its as though its an autonomous vehicle
4.	*CcR* ***(R)***:	what would it change if they reported it during the game
	CcR ***(L)***:	what would it change if they reported it during the game
	RcC ***(R)***:	it would cause a pause which is super enjoyable for the audience to watch
	RcC ***(L)***:	it would cause a pause which is super enjoyable for the audience to watch

Table 4: Attension weight distribution in reddit comment-reply pairs. Here *CcR* represents "Comment contextualized on Reply" whereas *RcC* represents "Reply contextualized on Comment"; ***(R)*** & ***(L)*** represents forward & backward attention.

The attention scores generated by the attention mechanism makes the proposed model highly interpretable. Table 4 showcases the distribution of the attention scores over four sarcastic (correctly predicted by Bi-ISCA) comment-reply pairs from the SARC dataset. Not only the proposed model was correctly able to detect sarcasm in these pairs of sentences but was also able to correctly identify words responsible for contextual, explicit, or implicit incongruity which invokes sarcasm.

For example in Pair 1, Bi-ISCA correctly identified explicitly incongruous words like *"amazing"* and *"force"* in the reply sentence which were responsible for the sarcastic nature of the reply. Interestingly the word "traumatized" in the parent

comment also had a high attention weight value, which shows that the proposed attention mechanism was able to learn the contextual incongruity between the opposite sentiment words like "traumatized" & "amazing" in the comment-reply pair. Pair 2 demonstrates the model's ability to capture words responsible for invoking sarcasm by making sentences implicitly incongruous. Sarcasm due to implicit incongruity is usually the toughest to perceive. Despite this, Bi-ISCA was able to give high attention weights to words like "announces" and "crashes & security holes". Not only this, but the proposed intra-sentence attention mechanism was also able to learn a link between "microsoft" and "m" (slang for microsoft) without having any prior knowledge related to slangs. Pair 3 is also an example of an explicitly and contextually incongruous comment-reply pair, where the model was successfully able to capture opposite sentiment words & phrases like "blind drunk", "cautious" and "behind the wheel" that made the reply sarcastic in nature. Pair 4 is an example of sarcasm due to implicit incongruity between the words, "pause" & "watch", and contextual incongruity simultaneously between "reported" & "enjoyable", both of which were successfully captured by Bi-BISCA.

7 Conclusion

In this paper, we introduce a novel Bi-directional Inter-Sentence Attention mechanism based model (Bi-ISCA) for detecting sarcasm. The proposed model not only was able to capture both intra and inter-sentence dependencies but was able to achieve state-of-the-art results in detecting sarcasm in the user-generated short text using only the conversational context. Further investigation of attention maps illustrated Bi-ISCA's ability to capture explicitly, implicitly, and contextually incongruous words & phrases responsible for invoking sarcasm. The success of the proposed model is achieved due to the use of character-based embeddings that takes care of slang/shortened & out of vocabulary words, Bi-LSTMs that captures intra-sentence dependencies between words in the same sentence, and Bi-ISCA that captures inter-sentence dependencies between words of different sentences.

References

[Amir *et al.*, 2016] Silvio Amir, Byron C Wallace, Hao Lyu, Paula Carvalho, and Silva Mário J. Modelling context with user embeddings for sarcasm detection in social media. *Proceedings of the Conference on Natural Language Learning (CoNLL)*, 2016.

[Bamman and Smith, 2015] David Bamman and Noah A Smith. Contextualized sarcasm detection on twitter. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[Barbieri *et al.*, 2014] Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. Modelling sarcasm in twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[Baruah *et al.*, 2020] Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. Context-aware sarcasm detection using BERT. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 83–87, Online, July 2020. Association for Computational Linguistics.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Dong *et al.*, 2020] Xiangjue Dong, Changmao Li, and Jinho D. Choi. Transformer-based context-aware sarcasm detection in conversation threads from social media. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 276–280, Online, July 2020. Association for Computational Linguistics.

[Eisterhold *et al.*, 2006] Jodi Eisterhold, Salvatore Attardo, and Diana Boxer. Reactions to irony in discourse: evidence for the least disruption principle. *Journal of Pragmatics*, 38(8):1239 – 1256, 2006. Focus-on Issue: Discourse and Conversation.

[Farías *et al.*, 2016] Delia Irazú Hernaundefineddez Farías, Viviana Patti, and Paolo Rosso. Irony detection in twitter: The role of affective content. *ACM Trans. Internet Technol.*, 16(3), July 2016.

[Ghaeini *et al.*, 2018] Reza Ghaeini, Xiaoli Z. Fern, and Prasad Tadepalli. Attentional multi-reading sarcasm detection. *CoRR*, abs/1809.03051, 2018.

[Ghosh and Veale, 2016] Aniruddha Ghosh and Tony Veale. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California, June 2016. Association for Computational Linguistics.

[Ghosh and Veale, 2017] Aniruddha Ghosh and Tony Veale. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[Ghosh *et al.*, 2020] Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online, July 2020. Association for Computational Linguistics.

[Hazarika *et al.*, 2018] Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Jaiswal, 2020] Nikhil Jaiswal. Neural sarcasm detection using conversation context. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 77–82, Online, July 2020. Association for Computational Linguistics.

[Joshi *et al.*, 2015] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China, July 2015. Association for Computational Linguistics.

[Khattri *et al.*, 2015] Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 25–30, 2015.

[Khodak *et al.*, 2018] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. In *Proceedings of the Linguistic Resource and Evaluation Conference (LREC)*, 2018.

[Kumar and Anand, 2020] Amardeep Kumar and Vivek Anand. Transformers on sarcasm detection with context. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 88–92, Online, July 2020. Association for Computational Linguistics.

[Kumar *et al.*, 2020] A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati, and L. B. M. Neti. Sarcasm detection using multi-head attention based bidirectional lstm. *IEEE Access*, 8:6388–6397, 2020.

[Lecun *et al.*, 1998] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lee *et al.*, 2020] Hankyol Lee, Youngjae Yu, and Gunhee Kim. Augmenting data for sarcasm detection with unlabeled conversation context. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 12–17, Online, July 2020. Association for Computational Linguistics.

[Liebrecht *et al.*, 2013] Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[Poria *et al.*, 2016] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1601–1612, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[Rajadesingan *et al.*, 2015] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 97–106, New York, NY, USA, 2015. Association for Computing Machinery.

[Reyes *et al.*, 2013] Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268, 2013.

[Riloff *et al.*, 2013] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 704–714. ACL, 2013.

[Schuster and Paliwal, 1997] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[Srivastava *et al.*, 2020] Himani Srivastava, Vaibhav Varshney, Surabhi Kumari, and Saurabh Srivastava. A novel hierarchical BERT architecture for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97, Online, July 2020. Association for Computational Linguistics.

[Tay *et al.*, 2018] Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[Tsur *et al.*, 2010] Oren Tsur, Dmitry Davidov, and Ari Rappoport. Icwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *fourth international AAAI conference on weblogs and social media*, 2010.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

[Wilson, 2006] Deirdre Wilson. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722 – 1743, 2006. Language in Mind: A Tribute to Neil Smith on the Occasion of his Retirement.

Modelling and Reasoning for Indirect Sensing over Discrete-time via Markov Logic Networks

Athanasios Tsitsipas* , **Lutz Schubert**
Ulm University, Germany
{firstname, surname}@uni-ulm.de

Abstract

With the always increasing availability of sensor devices, there is constant unseen monitoring of our environment. A physical activity has an impact on more sensor modalities than we could imagine. It is so vivid that distinctive patterns in the data look almost interpretable. Such knowledge, which is innate to humans, ought to be encoded and reason upon declaratively. We demonstrate the power of Markov Logic Networks for encoding uncertain knowledge to discover interesting situations from the observed evidence. We formally relate distinguishable patterns from the sensor data with knowledge about the environment and generate a rule basis for verifying and explaining occurred phenomena. We demonstrate an implementation on a real dataset and present our results.

1 Introduction

With the always-changing physical environments, uncertainty and incompleteness are innate in them. Context-aware pervasive systems have been the centre of research regarding approaches to modelling uncertain contextual information and reasoning upon it [Bettini *et al.*, 2010]; moving from low-level contextual data (i. e., sensors) to higher-level contextual information, where it is most commonly referred to as "situation" [Dey, 2001; Gellersen *et al.*, 2002]. Setting up systems to observe an environment includes deploying probes (e. g., sensors) tailored to specific situations. Today, such efforts fell under the terms "internet of things" and "smart homes". Many situations are worth identifying using sensors in a single room, ranging from "is someone present" to "water boiling". Considering an entire home, we may end up with hundreds of such situations. An office building could have thousands, increasing dedicated sensors to cover all the above situations, driving higher economic and maintenance costs.

A compelling method in such deployments is to use indirect sensing, which is employed when the property in need (e. g., a situation) is not attainable to direct sense, either due to sensor malfunctions, connectivity issues or energy loss. In the literature, indirect sensing is interwoven with remote sensing or sensing from afar [Zhang *et al.*, 2019]. In our study, we translate indirect sensing to a cooperative model of sensor fusion [Durrant-Whyte, 1990], where surrounding heterogeneous sensors capture different aspects of the same phenomenon (i. e., activity[1]). Activity is often described by a specific temporal organisation of low-level sensor data, or as we call it, a "dimensional footprint"(DF). The low-level sensor data in a DF are the primary source of information used as evidence to understand and recognise the observed situation. Such techniques following a bottom-up approach to recognising situations are well-established in the area of context-aware pervasive computing [Schmidt, 2003]. Dealing with a concept as the DF requires handling both *uncertainty* and the *relational organisation*. Existing approaches for an indirect sensing task typically fail to capture such aspects at the same time.

For the mechanics of an indirect sensing task, recent research targets data analysis techniques employing machine learning to train complex models labelling the property they want to infer from the data. For example, in [Laput *et al.*, 2017] the authors train Support Vector Machine (SVM) models, in an automatic learning mode à la "programming by demonstration" [Dey *et al.*, 2004; Hartmann *et al.*, 2007], with raw sensor data while performing the activity of interest. The major limitation of such systems is that they use representations that are not relatable to humans. In addition, they do not support explicit encoding of knowledge about the environment. Background knowledge (e. g., contextual, domain or commonsense) may describe situations absent in training data or challenging to grasp and annotate. In addition, apart from the definition of knowledge, the occurred observables (i. e., events) in sensor data may be uncertain, as much as the manifestations of knowledge are (i. e., rules) in an analytical reasoning process.

We address these limitations by choosing a probabilistic logic-based approach using an amalgam of Event Calculus (EC) [Kowalski and Sergot, 1989] and Markov Logic Network (MLN) [Richardson and Domingos, 2006] to model uncertain knowledge about the relational manifestations of different and heterogeneous sensors reasoning to infer interesting situations. EC drives the modelling task by a set of meta-

*Contact Author

[1] A situation, in that case, is the *state* of activity.

rules that encode the interaction between the sensor events and their effects over discrete time. One of the exciting properties of EC is that a situation of interest persists over time unless it gets interrupted by the occurrence of other events. On the other hand, MLN combines first-order logic and concepts from probability theory to tackle uncertainty, which has received considerable attention in recent years with applications in video activity analysis [Cheng *et al.*, 2014], maritime surveillance [Snidaro *et al.*, 2015], music analysis [Papadopoulos and Tzanetakis, 2016] and others. Our goal is to design a reasoning mode for indirect sensing that handles uncertainty and uses interpretable representations from data. To this end, we make the following contributions: (1) We model existing sensor data into interpretable symbolic representations as elements in a narrative on a running scenario (cf. Section 2.2), (2) design a knowledge base (KB) within MLN for supporting indirect sensing while emulating commonsense reasoning, (3) evaluate the realisation of the approach using an open-source implementation of MLN, (4) demonstrate how the probability of an occurred situation changes over time while using different combinations of sensors.

Section 2 provides the terminology used in this document, including the running example and background information on Event Calculus and Markov Logic Networks. This leads to Section 3 where we introduce the concept of DF and how to model it. In Section ,4 we elaborate on MLN definitions, while in Section 5, we present the results and experiments. Section 6 provides a brief related work around the topic of event modelling and recognition. In Section 7, we summarise the main contributions and discuss details, including future work.

2 Preliminaries

2.1 Terminology

The Oxford English Dictionary gives a general definition for an *event* as "*a thing that happens or takes place, especially one of importance*". In our context, a "*thing*" is represented by a (sensor) data pattern. "*Importance*" matches the (subjective) interest in finding an explanation for this pattern. Many researchers try to use the term event in their way, depending on the context and the investigated environment, even though the definition of the word event remains the same.

We assume that an (interesting) event occurred on identifying a visible change in the sensor data. The identification involves a pre-processing step using some pattern extraction techniques [Patel *et al.*, 2002; Lin *et al.*, 2003; Yeh *et al.*, 2016]. Therefore, the timestamps for the respective pattern represents the event's *temporality*. This work clarifies a *time point* and a *series of time points* (exhibiting the concept of *duration*) bounded by a predefined window value. For example, the increase in the temperature readings is an interesting event and reflects the development of sensing data (temperature) over time. Therefore, a representation should semantically annotate an event's time point.

Interpreting symbols as representations of objects is a proxy to describe something instead of the actual thing. For example, if something is an ambient "high" temperature[2], that temperature does not reside in our heads when we think of it. The "it" of the temperature is a representation of the actual natural environmental property. This representation of something is an entity that transmits to us the idea of the real something. Perhaps we think of our discomfort or imaging ourselves reacting to this phenomenon (e. g., sweating) to represent the high ambient temperature. Alternatively, we use the colour red accompanied by the temperature degree.

An event *representation* in our work is a lexical word embedded in a "sentence" among other additional contextual words, which we understand. Therefore, the development of sensing data over time (i. e., a time series) is wrapped in a word that best describes its nature (e. g., data pattern). The event representation has two lexical parts. The one part is the *trend* of the pattern, and the other one is the *type of the pattern*. The trend of a pattern is represented by the words *upward* or *downward*. The patterns we may derive in the sensor readings could resemble a shape currently named *shapeoid*. For the sake of presentation, the lexical *shapeoids* are the following:

ANGLE A gradual, continuous line with an increasing (upward) or a decreasing (downward) trend in the sensor readings.

HOP A stage shift in the sensor readings, where the data have an apparent difference between two consecutive recognition time points (e. g., binary sensor values).

HORN This pattern is a transient increase or decrease in the sensor readings curve.

FLAT A horizontal line in the data, with either unchangeable values in the pattern duration or minimal changes.

We extract the *shapeoids* using the Symbolic Aggregate Approximation (SAX) technique. Many time series representation alternatives exist, but most of them result in a downsampled real-valued representation. In contrast, SAX boils down to a symbolic discretised form of the time series, which is abstract enough to extract the *shapeoids* generally. The paper's focus is not to describe how to obtain the proposed patterns from the sensor data but to put forward a concept of using temporal organisations of such representations to reason in a robust and declarative way.

2.2 Running Example

In Figure 1, we illustrate the activity of opening and closing a window and its impact (i. e., their DF) on five surrounding sensor types that happen to be in the same room. Later in the paper (cf. Section 3.3), we will showcase the extracted *shapeoids* from the raw data, which put forward a sufficient abstraction, serving as an input for a reasoning task.

The data are from a real-world public dataset [Birnbach *et al.*, 2019], where the authors collected sensor data while performing different activities. The data timeline spawns over two minutes, sufficient for demonstrating the essence of our approach.

[2]We use a threshold-based term to describe the comfort level for a human to endure.

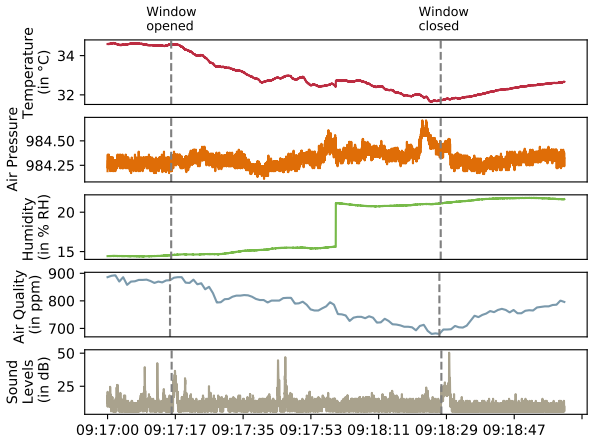


Figure 1: An example of how the activity of opening/closing a window affects the listed surrounding sensors.

Predicate	Meaning
$\mathsf{Happens}(e,\ t)$	Event e happens at time t
$\mathsf{HoldsAt}(f,\ t)$	Fluent f holds at time t
$\mathsf{InitiatedAt}(f,\ t)$	Fluent f is initiated at time t
$\mathsf{TerminatedAt}(f,\ t)$	Fluent f is terminated at time t
Axioms	
$\mathsf{HoldsAt}\,(f,\ t+1) \Leftarrow \mathsf{InitiatedAt}\,(f,\ t)$	$\mathsf{HoldsAt}\,(f,\ t+1) \Leftarrow \mathsf{HoldsAt}\,(f,\ t) \wedge \neg\,\mathsf{TerminatedAt}\,(f,\ t)$
$\neg\,\mathsf{HoldsAt}\,(f,\ t+1) \Leftarrow \mathsf{TerminatedAt}\,(f,\ t)$	$\neg\,\mathsf{HoldsAt}\,(f,\ t+1) \Leftarrow \neg\,\mathsf{HoldsAt}\,(f,\ t) \wedge \neg\,\mathsf{InitiatedAt}\,(f,\ t)$

Table 1: The core predicates and domain-independent axioms of the EC dialect, MLN-EC.

2.3 Event Calculus

Representing and reasoning about actions and temporally-scoped relations has been a critical research topic in the area of Knowledge Representation and Reasoning (KRR) since the 60s [Shoham and McDermott, 1988]. Since then, various approaches have been proposed to overcome the Frame Problem in classical Artificial Intelligence (AI) [McCarthy and Hayes, 1981; Shanahan, 2006]; the challenge of representing the effects of actions. Among them, EC, which Kowalski and Sergot have initially proposed in 1986 [Kowalski and Sergot, 1989], is a system for reasoning about events (or actions) and their effects in the scope of Logic Programming. It comprises excellent expressiveness with intuitive and readable representations, making it feasible to extend reasoning. It is an adequate tool to fit domain knowledge representing how an entity progresses in time using events. It has found applications ranging from the scope of robotics [Russel *et al.*, 2013], game design [Nelson and Mateas, 2008] and commonsense reasoning [Shanahan, 2004; Mueller, 2014] to name a few.

From a technical point, the core ontology of the EC involves *events*, *fluents* and *time points*. The continuum of time is linear, and integers or real numbers represent the time points. A *fluent* can be whatever whose value is subject to change over time. At the occurrence of an *event*, it may change the value of a fluent. This could be a quantity, such as "the temperature in the room", whose value varies in numbers, or a proposition, such as "the window is open", whose truth state changes from time to time. In EC, the core axioms are domain-independent and define whether a fluent holds or not at a particular time point. In addition, these axioms can capture what is known as the common sense *law of inertia*; formal logic is a way of declaring that an event is assumed not to change a given property of a fluent *unless* there is evidence to the contrary [Shanahan and others, 1997].

We use a simplified version of EC (named MLN-EC), based on a discrete-time reworking of EC [Mueller, 2008], which was proven to work in a probabilistic setting [Skarlatidis *et al.*, 2015]. Other dialects may have additional restrictions (e. g., complex time quantification) that hinder the realisation of the approach. For more information, we point the reader to this paper [Mueller, 2004]. The basic predicates and the domain-independent axioms are presented in Table 1. One can read the upper line of two axioms from left to right: (1) a fluent f holds at time t if it was initiated at a previous time point, and (2) that the fluent f continues to hold, providing it was not previously terminated. The domain-dependent predicates $\mathsf{initiatedAt}/2$ and $\mathsf{terminatedAt}/2$ are expressed in an application-specific manner guiding the logic behind the occurrence of events and some contextual constraints. One example of a common rule for $\mathsf{initiatedAt}/2$ is:

$$\mathsf{InitiatedAt}\,(f,\ t) \Leftarrow \mathsf{Happens}\,(e,\ t) \wedge \mathit{Constraints}[t] \tag{1}$$

The above definition states that a fluent f is initiated at time t if an event e happens, and some optional constraints depend on the domain. EC supports default reasoning via circumscription, representing that the fluent continues to persist unless other events happen. Therefore, in our definition of the event narrative, we assume these are the only events that occurred.

2.4 Markov Logic Networks

A Markov Logic Network (MLN) amalgam of a Markov Network (aka. Markov Random Field) and a first-order logic KB [Richardson and Domingos, 2006]. Specifically, it softens the constraints posed by the formulas with weights that support (positive weights) or penalise (negative weights) worlds in which they are satisfied. As opposed to classical logic, all the statements are hard constraints (i. e., preserving truthfulness).

The formulas, being first-order logic objects [Genesereth and Nilsson, 1987], are constructed using four symbols: *constants*, *variables*, *functions* and *predicates*. Predicates and constants start with an upper-case letter, whereas the functions and variables have lower-case letters. The variables are quantifiable over the given domain (e. g., type={Temperature, Humidity}). The *constants* are objects in the respective domain (e. g., sensor types: Temperature, Air Quality, Microphone etc.). *Variables* are ranges over the objects of the

domain. The *functions* (e. g., downwardAngleTemp) represent actual mappings from a single object to a value or another object. Finally, the *predicate* symbols represent relations among objects associated with truth values(e. g., Happens(DownwardAngle_Temp,4)).

A KB in MLN consists of both hard- and soft-constrained formulas. Hard constraints (clauses with infinite weight) are interwoven with unequivocal knowledge. Therefore, an acceptable world fulfils all of the hard constraints. By contrast, the soft constraints are related to the imperfect knowledge of the domain, which can be falsified in the world's existence in discourse. This means that when a world violates a formula, it is less probable but not impossible.

Formally, a MLN is a set of pairs (F_i, w_i), where F_i is a first-order logic formula and w_i is a real numbered weight. The KB L, with the weighted formulas together with a finite set of constants $C = \{c_1, c_2, \ldots, c_{|C|}\}$, defines a ground Markov Network $M_{L,C}$ as follows [Richardson and Domingos, 2006]:

- $M_{L,C}$ has one binary node for each possible grounding of each predicate in L. The value of the node is 1 if the grounded atom is true and 0 otherwise.
- $M_{L,C}$ contains one feature for each possible grounding of each formula F_i in L. The value of this feature is 1 if the formula is true and 0 otherwise. The weight of the feature is the w_i associated with F_i in L.

An MLN is a template for constructing Markov networks: it will produce different networks given different constants. The *grounding* process is the replacement of variables with a constant in their domain. The nodes of $M_{L,C}$ correspond to all ground atoms that can be generated by grounding a formula F_i in L, with constants of C. Thus there is an edge between two nodes of $M_{L,C}$ iff the corresponding ground predicates are conditionally dependant on a grounding of a formula F_i in L. A possible world from the MLN must satisfy all of the hard-constrained formulas and be proportional to the exponential sum of the weights of the soft-constrained formulas satisfied in this world (cf. Equation 2). Hence, a MLN defines a log-linear probability distribution over Herbrand interpretations(i. e., possible worlds).

In an indirect sensing task context, we know *a priori* that we will have two kinds of predicates; the evidence variable X, containing the narrative of real-time input events, translated with the Happens predicates of EC, and the set of query HoldsAt predicates Y, as well as other groundings of "hidden" predicates (i. e., neither query nor evidence); in EC these are the InitiatedAt and TerminatedAt predicates. Finally, the conditional likelihood of Y given X is defined as follows [Singla and Domingos, 2005]:

$$P(y \mid x) = \frac{1}{Z_x} \exp\left(\sum_{i \in F_Y} w_i n_i(x, y)\right) \qquad (2)$$

$x \in X$ and $y \in Y$ represent the possible assignment of the evidence set X and the query set Y, respectively. F_Y is the set of all MLN clauses produced from the KB L and the finite set of constants C. The $n_i(x, y)$ is the number of true groundings of the i-th clause involving the query atoms y given the evidence atoms x. Finally, Z_x is a partition function that normalises for all the possible assignments of x.

Equation 2 shows the probability distribution of the set of query variables conditioned over the set of observations. By modelling the conditional probability directly, the model remains agnostic about potential dependencies between the variables in X, and any factors that depend on X are eliminated. Instead, the model makes conditional independence assumptions among the Y and assumptions on its inherent structure with dependencies of Y on X. Therefore, in such a way, the number of the possible words is constrained [Singla and Domingos, 2005; Sutton and McCallum, 2006] and the inference is much more efficient. However, calculating exactly the formula might become intractable even for a small domain. Consequently, other approximate inference methods are preferred.

Originally, the authors in [Richardson and Domingos, 2006] propose to use Gibbs sampling to perform inference, but they found out that the sampling breaks down when the KB has deterministic dependencies[3] [Poon and Domingos, 2006; Domingos and Lowd, 2009]. The authors proposed another Markov Chain Monte Carlo method called MC-SAT [Poon and Domingos, 2006] based on satisfiability with slice-sampling. Another type of inference is the *Maximum A Posteriori* (MAP) which described the problem of finding the most probable state of the world given some evidence, which reduces to find the truth assignment that maximises the sum of weights of satisfied clauses (i. e., $\underset{y}{\operatorname{argmax}} p(y \mid x)$).

The problem is generally NP-hard, but both exact and approximate satisfiability solvers exist [Domingos and Lowd, 2009]. In our experiments, we run approximate inference using the MC-SAT algorithm.

3 Modelling a DF

An activity affects various fundamental environmental properties, such as speed, pressure, temperature, luminosity, etc. Surrounding sensors may capture the various changes (forming the activity's DF), which depends on different contextual information, such as their proximity from the occurred phenomenon and their type (cf. Section 3.1). In addition, a sensor may observe ambient values (e. g., temperature) or require manual intervention to observe a change (e. g., separating the two magnetic elements of a contact sensor) (cf. Section 3.2).

This *"change"* (i. e., the forming pattern) is the "interesting event" we want to focus on. This observed change mostly stays unobserved. Thus, the emitted DF indicates its occurrence. In addition, its state is a continuous value in time, which is tracked under the definition of the *"fluent"*. With no sensor modality to identify the occurrence of an activity, due to its unavailability at the given time, or by simply stating that there does not exist any direct one, we account its DF as a *space* with equivalent options that "indirectly" account for the same activity.

Our work uses commonsense knowledge (CK) to characterise how activity affects its environment. From the running

[3] They are formed from hard-constrained formulas in the KB.

example in Section 2.2, some distinct data patterns exist, almost as recognisable to the human eye where one may exercise a hypothesis against the data. We consider that a data-processing step is viable to extract such patterns, but it is out of the scope of the current paper. The abstracted representations (cf. Section 2.1) from low-level sensor data reflect their organisations in shapes and trends (e. g., an increasing angle in the sensor data). Therefore, one with a naive knowledge of physics can make hypotheses about the occurrence of an activity using the abstractions from sensor data as evidence (cf. Section 3.3).

3.1 Contextual Constraints

Sensors are interfaces that serve as occurrence indicators for various monitored situations. The sensor numbers could increase accordingly as their numbers increase, making the instrumentation, deployment and maintenance cumbersome tasks. A sensor primarily measures an environmental change as accurate as possible, varying between the different manufacturers. Selecting a sensor to monitor a situation ought to obey some criteria, which formulate the sensing fidelity of its output. In this paper, we propose the following criteria:

Type There exist different vendors for various sensors. Nonetheless, the type of sensor is of key importance. There is no doubt that different manufacturers may offer a better sensor device, affecting accuracy. Semantically, the sensor type determines if the sensor participates in the verification process, not its model.

Location The location is another important aspect of determining the credibility of the sensor output. Either the physical location or the position of the sensor in the space should affect the decision of selecting any sensor of a given type in a location (e. g., a room).

As discussed later in the paper, the above criteria are minimal constraints for a sensor to participate in reasoning. However, the sensors have a fundamental high-level classification, making the *shapeoid* extraction from their data clearer and focused.

3.2 Sensor Classification

A sensor is an interface between the physical and the digital world. The raw sensor data rarely matches human semantics, but the representations of patterns in them are. The kind of sensor classification the paper foresees, bases on the nature of the resulting sensor data, is as follows:

Binary sensors restrict their result to two possible values. Usually, the values resemble the category itself (i. e., being binary); thus, *one* and *zero*. Furthermore, depending on the context[4] the result may take values from it. For example, the output of a physical switch is "on" or "off", the result of a motion sensor may be "present" or "not present", and so on. The suitable data patterns for the binary sensors are the HOP and FLAT representations.

Numerical sensors are almost every sensor with an arithmetic output in the set of real numbers $\mathbb{R}$. Some examples of quantifiable sensors are an accelerometer, humidity, temperature, pressure sensor etc.. Accordingly, the data patterns, which we found in the raw sensor data, are those of ANGLE, HORN and FLAT.

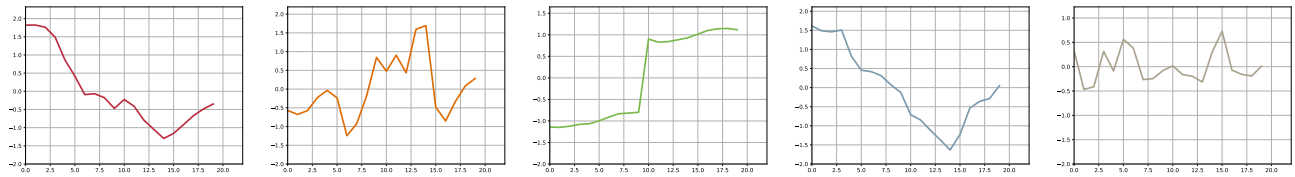
One could say that a binary sensor is a subset of numerical sensors. However, we make the distinction explicit, as the binary sensors are semantically a practical standalone class. In the running example, we use numerical sensors. The sensor data's available observations (i.e., shapeoids) are the simple events with their respective time point in the focused bounded time window. We represent them with the Happens predicate, where finally a collection of such predicates form the so-called "narrative" in EC.

3.3 The Narrative of Events in EC

An event "just" *happens*, with an accompanied discrete time point to keep a reference in the timeline. The chosen representation of it, according to the dialect of EC, is the predicate $\mathsf{Happens}(e, t)$. Time t can be quantified over the spectrum of integers, exhibiting coherence among the occurred events. The events in the sensing timeline are the formed *shapeoids*, and by using lexical words for the symbolic representation, the intuition behind them is human-readable (e. g., *downward* ANGLE). For example, in Figure 1, the two activities of opening and closing the window produce an impact in the five surrounding sensors. We observe that around the time of opening the window, distinct patterns are forming. Figure 2 contains in separate graphs a more clear view of the data in Figure 1, after performing a dimensionality reduction step (e. g., Piecewise Aggregate Approximation (PAA) [Ding *et al.*, 2008]). The patterns were extracted empirically, resembling the proposed lexical *shapeoids* (cf. Section 2.1):

$$
\begin{aligned}
&\ldots\\
&\mathsf{Happens(Flat\_Mic,3)}\\
&\mathsf{Happens(Flat\_Hum,3)}\\
&\mathsf{Happens(DownwardAngle\_Temp,4)}\\
&\mathsf{Happens(DownwardAngle\_Aq,4)}\\
&\mathsf{Happens(UpwardHorn\_Mic,4)}\\
&\ldots\\
&\mathsf{Happens(UpwardHorn\_Temp,11)}\\
&\mathsf{Happens(UpwardAngle\_Hum,11)}\\
&\mathsf{Happens(Flat\_Mic,11)}\\
&\ldots\\
&\mathsf{Happens(DownwardAngle\_Temp,14)}\\
&\mathsf{Happens(DownwardAngle\_Pres,14)}\\
&\mathsf{Happens(Flat\_Hum,15)}\\
&\mathsf{Happens(UpwardHorn\_Mic,15)}\\
&\mathsf{Happens(UpwardAngle\_Temp),15)}\\
&\ldots
\end{aligned}
\tag{3}
$$

[4] *Context is any information that one can use to characterise the situation of an entity. An entity is a person, place, or object considered relevant to the interaction between a user and an application, including the user and application themselves* [Dey, 2000]



(a) Temperature (b) Air Pressure (c) Humidity (d) Air Quality (e) Microphone

Figure 2: The z-normalised sensor data (in 20 data points) from Figure 1, after a dimensionality reduction step.

4 Probabilistic Indirect Sensing via MLN definitions

.

In the following, we elaborate on constructing the KB containing the representations of the sensor events, using contextual words in "sentences" that comply with the formalism of EC and are expressed in first-order logic.

4.1 Knowledge Base

For our purposes, the KB, or the so-called "theory", contains a few function definitions, predicate definitions, as well as the inertia laws axioms of EC[5] as seen in (2). We consider the observed patterns as a continuous narrative of Happens predicates (cf. (3)). InitiatedAt and TerminatedAt determine under which factors a fluent is initiated or terminated at a given time point, using the form in (1). Finally, the query predicate HoldsAt incorporate a possible quantification over the verification of a monitored situation (i. e., a fluent).

Table 2 shows a fragment of the KB and the associated weights. The formulas are converted to a clausal form during the grounding phase, also known as conjunctive normal form (CNF), a disjunction of literals. The next step is the replacement of the variables with the constants, which formulate grounded predicates. As such, the construction of the Markov Network consists of one binary node V for each possible grounding of each predicate. A world is an assignment of a truth value to each of these nodes.

The definition of the indirect sensing rules follow CK represented as a theory in MLN enacting it as part of commonsense reasoning (CR)[6]; the sort of reasoning people perform in daily life [Mueller, 2014], which is vague and uncertain. For example, the Table 2 contains two separate rules, which reflect an atomic instruction of the DF, using a temperature sensor and a microphone. For our purposes, we consider that the events in the narrative are the only one occurred.

A rise in the temperature readings, or a sudden spike in the sound pressure levels, could be anything in an open world, including the opening/closing of a door in a room. However, with the help of context, we may exercise the hypothesis that a temperature sensor *close* to the window could indicate its state. The hypothesis is asked in the form of a query, representing the probability for the situation of *an opened window* to be true for the given observations (i. e., ground truth). For example, if we require to encode an "opened door", we may include the same rule with a lower weight encoding our confidence for the result. Then, using the background knowledge that the sensors are closer to the window, we encode this with a higher weight value to the *opened window* rule. MLN has many learning algorithms [Richardson and Domingos, 2006] to determine the weight assignment; however, as we do not intend to select the absolute probabilities of a specific occurred situation, we opt for the most likely situation given the evidence.

4.2 Evidence

The evidence contains ground predicates (facts) (e. g., the narrative of events in Section 3.3) and optionally ground *function mappings*. A *function mapping* is a process of mapping a function to a unique identifier. For example, the first formula in Table 2 contains the function downwardAngleTemp(r). During the grounding phase, constants from the domain of the variable r substitute it[7]. Thus, a function mapping could be the following: DownwardAngle_Temp_LocA = downwardAngleTemp(LocationA). All the events of the grounded Happens predicates in Section 3.3 follow the same procedure for their function mappings.

5 Experiments and Results

In this section, we evaluate our approach in the domain of smart homes. As presented in Section 2.2, we use a publicly available dataset. The data timeline spawns over twelve consecutive full days. The dataset was in a zip format, which contains multiple comma-separated value (CSV) files with a total size of approximately 50 Gigabytes (GB)[8]. We selected one device close to the interest situation (i.e., close to the window). We extracted the relevant data points using the five sensors capturing the DF of opening/closing the room's window. We do not process the raw data points, but instead, we use the *shapeoids* from the data; their extraction was possible via our tool *Scotty*[9]. The total number of shapeoid events are 4393, where the ground truth events from the window contact sensor are 87.

[5]They should remain hard-constrained; otherwise, the recognition of the situation will converge to be uncertain up to the horizon of probability.

[6]CR is implemented as a valid (or approximately valid) inference [Davis, 2017] in MLN as part of the EC law of inertia.

[7]We assume a single room and its context is not reflected in the naming scheme of the function.

[8]The actual size of the raw data exceeds the 250 GB.

[9]This work is meant to be published in a forthcoming conference.

FOL formula	Weight
$\mathsf{InitiatedAt}\,(\mathsf{openedWindow}\,(r)\,,\,t) \Rightarrow \mathsf{Happens}\,(\mathsf{downwardAngleTemp}\,(r)\,,\,t) \wedge \mathsf{Happens}\,(\mathsf{flatTemp}\,(r)\,,\,t-1)$	2.1
$\mathsf{InitiatedAt}\,(\mathsf{openedWindow}\,(r)\,,\,t) \Rightarrow \mathsf{Happens}\,(\mathsf{upwardHornMic}\,(r)\,,\,t) \wedge \mathsf{Happens}\,(\mathsf{flatMic}\,(r)\,,\,t-1)$	0.2

Table 2: An excerpt of the first-order KB and the corresponding weights in the MLN.

Scenario	Description	Duration
S#1	Two sensors with weak and strong weights.	1 m 45 s
S#2	Three sensors with one weak and two strong weights.	1 m 9 s

Table 3: The described scenarios with their inference duration times.

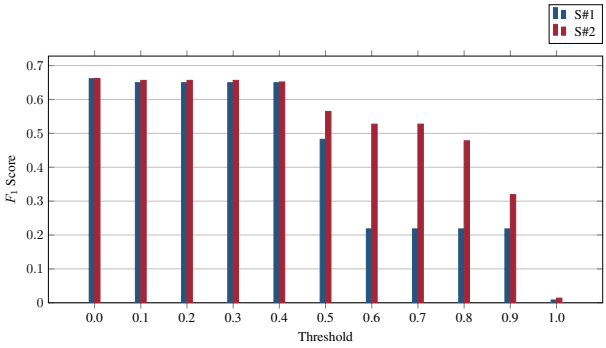


Figure 3: F_1 scores using various threshold values for the situation recognition of the opened window.

We put forward two scenarios (cf. Table 3), which contain rules for declaring the alternatives in recognising the situation of an opened/closed window. The purpose of the scenarios is to run the computation against the existing narrative with the discovered events but using different sensor compositions. Each recognition rule also contains a weight value, which was empirically assigned, as we consider them confidence values of the rule.

We implemented the KB and the narrative evidence file to demonstrate the approach's feasibility using an open-source implementation of Markov Logic Networks, named LoMRF [Skarlatidis and Michelioudakis, 2014]. Together with the domain-dependent rules for each scenario, the full KB and the evidence file are publicly available online[10], enabling the reproducible results. The KB, given the evidence, is transformed into a Markov Network of 26353 ground clauses and 13177 ground predicates. We run marginal inference from the developed MLN on vanilla runs without any interference from other processes. All the results are averaged over five runs with a corresponding standard deviation. The experiments are executed on a virtual machine(VM) running in a self-hosted data centre at the University of Ulm running on OpenStack under the series "Victoria". The VM runs with 8 cores (16 threads) and 16 GB of RAM.

[10]https://osf.io/n3ury/

Scenario	TP	TN	FP	FN	Precision	Recall	F_1
S#1	288	2039	174	1892	0.6234	0.1321	0.2180
S#2	1016	1554	659	1164	0.6066	0.4661	0.5271

Table 4: Performance results using the marginal inference and a threshold of 0.6.

In the experimental analysis, we present the results for the marginal inference in terms of F_1 score for a range of thresholds between 0.0 and 1.0. We consider the situation recognition task successful with a probability above the specified threshold. In Table 4, we present a snapshot of the performance using the threshold value 0.6 in terms of True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), Precision, Recall and F1 score.

The scenarios have a certain flavour. The basic intuition from the experiments is to showcase that we may use sensors that have an obscure interpretation (e. g., a spike in the microphone can be anything, even being next to the window) and sensors that act as a more direct verification step (e. g., air quality, temperature). We assume that the *shapeoid* events are the only ones that happen in the environment in focus. More alternative sentences may be encoded accordingly, using *shapeoids* of the humidity or the air pressure sensor. Based on the inertia laws of EC, the fluent start to hold at the time point *t+1*, and therefore the assignment to the next time point from the used pattern event in the narrative (3). In Figure ,3 the F_1 score is higher for the marginal inference in S#2 due to the additional strong sensor. The S#1, similar to S#2, contains a shapeoid in the microphone data (increasing horn), which matches both the fluent's initiation and termination rules. Hence, during the inference process, the probability always strives towards 0.5, which is regulated by another sensor in the rules (air quality sensor) with a higher weight value.

We note here that in a real setting, the verification of situation (i. e., the fluent) depends on whether the required observation is made (e. g., the shapeoid event from the temperature sensor), which may be a delayed effect of the activity itself - in other words: it takes some time until the open window affects the temperature sufficiently. In the experimental analysis, we calculate the performance measures strictly based on the time range of an opened window. Therefore the ground truth is the single point of reference for calculating the performance. The delay between the activity and its observable DF should be accounted for a more accurate timing prediction. We observe a considerable amount of FP, which indicates a plausible calculation of an opened window but with a certain recognition delay. Thus, we consider the F_1 scores in the scenarios to be slightly higher.

6 Related Work

Research in context modelling, context reasoning, and their unified view via various middleware systems is tremendous; for a recent survey, we point the reader to [Perera *et al.*, 2013]. In the paper, we focus more on a bottom-up approach to the recognition of occurred situations. We employ a probabilistic rule-based approach, using occurred sensor events as evidence for the reasoning task.

In [Liu *et al.*, 2017], the authors create a bottom-up hierarchical model using the raw sensor data as evidence while creating inference rules encoded in an MLN to recognise complex events. In order to create abstractions from the raw data, they use various thresholds per sensor type. In our approach, we use generic template abstractions which base on the data shapes and trends. The core contribution of their paper is the dynamic assignment of weights learned from a training dataset; we do not assume that the user has a training dataset to learn the weights from because we use them as confidence values for the inference rules. Finally, in our paper, we foresee scalability issues that may arise from the free variables in the MLN rules, which may drive the computation times to higher levels.

Considering our choice for a rule-based reasoning technique has a broad spectrum of applications to many domains, making it a commonly used technique [Perera *et al.*, 2013]. Another interesting technique, which bases on previously acquired knowledge, is case-based reasoning (CBR) [Aamodt and Plaza, 1994; Biswas *et al.*, 2014]. It offers solving mechanisms by adopting solutions that have been suggested to similar issues in the past. The authors in [Kofod-Petersen and Aamodt, 2003] use CBR to understand an occurred situation based on available contextual information. A case-based solution is not favourable in our case because collecting and maintaining previous cases is a cumbersome task. Our work does not require any previous known input from sensor observations and domain-dependent knowledge during the rule specification.

In the paper, we focus on finding alternatives for recognising a situation. Similarly, Loke [Loke, 2006] advocates that the situation *in_meeting_now* has different recognition ways based on contextual cues. The author follows an abductive treatment of the subject as we also do. In the forthcoming years, the author developed a formalism to represent compositions of sensors that can act on an understanding of their situations [Loke, 2016].

Finally, although sensing data contain implicit information, explicit domain knowledge is required for situation recognition. Many research works employ logic-based models for situation recognition in smart homes, such as the Event Calculus (EC) [Chen *et al.*, 2008]. Other works have also employed EC in activity recognition from video streams [Artikis *et al.*, 2014] and health monitoring [Falcionelli *et al.*, 2019]. However, it is unclear how they move from the raw data to the tagged symbolic representations in these systems.

7 Conclusion & Discussion

In the paper, we employed Markov Logic Networks for the modelling and reasoning over uncertain alternatives for the method of indirect sensing. We use the temporal formalism of EC as a "linchpin" for driving the reasoning about the sensing objects and creating observations for the occurrence of certain situations (e. g., "is the window open"). The concept of the DF allows using different sensor setups to monitor the same situation(s). In other words, it is parallel to interpreting the given evidence (e. g., sensor data) for finding the most likely explanation, which created the DF. As such, we declare these logical "inference" sentences in a human-readable form of reasoning that incorporates commonsense logic.

Due to the nature of environmental situations, the interpretation (i. e., evaluation) of such sentences depends on the full context. For example, the same sentence in Table 2 might not apply if the weather outside is warmer than the sensor's environment. In this case, the temperature may not decrease but stay the same or even increase. Therefore, one will never evaluate the according to sentence to true. Instead, a fallback to another sensor is needed. Nevertheless, the approach defends the redundancy, or alternatives, in detecting the desired situation, considering that we usually use direct means for sensing (e. g., use a contact sensor to detect if the door is open).

The lack of sensors to capture the whole DF of activity leads to an incomplete "view of the world". The question thereby is, which physical effects are of specific relevance for interpreting an event and omitted. These conditions may vary enormously between different events, e. g., a person speaking or the sun rising both have other effects on the environment and thus (to a degree) require various sensors for interpretation, but also both could be observed using additional information: sound, visual, temperature, time etc.

Concerning the employed method of MLNs, there is an issue using predicates with free variables in the body of a rule; during the grounding phase, it creates a disjunction of the cartesian grounded conjunction of the formulas, translating these variables to existentially quantified leading to a possible combinatorial explosion. We consider any additional constraint in a domain-dependent rule should contain as variables only the time t and the location r. Any knowledge engineer should follow this and remove any existentially quantified variables, using the technique of skolemisation [Broeck *et al.*, 2013], overcoming this limitation for the solution's scalability.

Finally, the observed data patterns may also result from multiple overlapping activities challenging to separate, such as speaking in traffic, leading to uncertainty about the interpretation. As future work, we want to overcome the limitations of MLN concerning the free variables in the rules and concentrate on a dynamic ecosystem that realises the proposed work.

Acknowledgments

This work was partially funded by the Federal Ministry of Education and Research (BMBF) of Germany under Grant No. 01IS18072.

References

[Aamodt and Plaza, 1994] Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1):39–59, 1994.

[Artikis *et al.*, 2014] Alexander Artikis, Marek Sergot, and Georgios Paliouras. An event calculus for event recognition. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):895–908, 2014.

[Bettini *et al.*, 2010] Claudio Bettini, Oliver Brdiczka, Karen Henricksen, Jadwiga Indulska, Daniela Nicklas, Anand Ranganathan, and Daniele Riboni. A survey of context modelling and reasoning techniques. *Pervasive and mobile computing*, 6(2):161–180, 2010.

[Birnbach *et al.*, 2019] Simon Birnbach, Simon Eberz, and Ivan Martinovic. Peeves: Physical event verification in smart homes. In *Proceedings of the 2019 ACM Conference on Computer and Communications Security*. ACM, 2019.

[Biswas *et al.*, 2014] Saroj K Biswas, Nidul Sinha, and Biswajit Purkayastha. A review on fundamentals of case-based reasoning and its recent application in different domains. *International Journal of Advanced Intelligence Paradigms*, 6(3):235–254, 2014.

[Broeck *et al.*, 2013] Guy Van den Broeck, Wannes Meert, and Adnan Darwiche. Skolemization for weighted first-order model counting. *arXiv preprint arXiv:1312.5378*, 2013.

[Chen *et al.*, 2008] Liming Chen, Chris Nugent, Maurice Mulvenna, Dewar Finlay, Xin Hong, and Michael Poland. Using event calculus for behaviour reasoning and assistance in a smart home. In *International Conference on Smart Homes and Health Telematics*, pages 81–89. Springer, 2008.

[Cheng *et al.*, 2014] Guangchun Cheng, Yiwen Wan, Bill P Buckles, and Yan Huang. An introduction to markov logic networks and application in video activity analysis. In *Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2014.

[Davis, 2017] Ernest Davis. Logical formalizations of commonsense reasoning: a survey. *Journal of Artificial Intelligence Research*, 59:651–723, 2017.

[Dey *et al.*, 2004] Anind K Dey, Raffay Hamid, Chris Beckmann, Ian Li, and Daniel Hsu. a cappella: programming by demonstration of context-aware applications. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 33–40, 2004.

[Dey, 2000] Anind Kumar Dey. *Providing Architectural Support for Building Context-Aware Applications*. PhD thesis, Georgia Institute of Technology, USA, 2000. AAI9994400.

[Dey, 2001] Anind K Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.

[Ding *et al.*, 2008] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

[Domingos and Lowd, 2009] Pedro Domingos and Daniel Lowd. Markov logic: An interface layer for artificial intelligence. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–155, 2009.

[Durrant-Whyte, 1990] Hugh F Durrant-Whyte. Sensor models and multisensor integration. In *Autonomous robot vehicles*, pages 73–89. Springer, 1990.

[Falcionelli *et al.*, 2019] Nicola Falcionelli, Paolo Sernani, Albert Brugués, Dagmawi Neway Mekuria, Davide Calvaresi, Michael Schumacher, Aldo Franco Dragoni, and Stefano Bromuri. Indexing the event calculus: towards practical human-readable personal health systems. *Artificial intelligence in medicine*, 96:154–166, 2019.

[Gellersen *et al.*, 2002] Hans W Gellersen, Albrecht Schmidt, and Michael Beigl. Multi-sensor context-awareness in mobile devices and smart artifacts. *Mobile Networks and Applications*, 7(5):341–351, 2002.

[Genesereth and Nilsson, 1987] Michael R Genesereth and Nils J Nilsson. *Logical foundations of artificial intelligence*. Morgan Kaufmann, 1987.

[Hartmann *et al.*, 2007] Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R Klemmer. Authoring sensor-based interactions by demonstration with direct manipulation and pattern recognition. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–154, 2007.

[Kofod-Petersen and Aamodt, 2003] Anders Kofod-Petersen and Agnar Aamodt. Case-based situation assessment in a mobile context-aware system. In *Artificial Intelligence in Mobile Systems*, pages 41–49, 2003.

[Kowalski and Sergot, 1989] Robert Kowalski and Marek Sergot. A logic-based calculus of events. In *Foundations of knowledge base management*, pages 23–55. Springer, 1989.

[Laput *et al.*, 2017] Gierad Laput, Yang Zhang, and Chris Harrison. Synthetic sensors: Towards general-purpose sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3986–3999, 2017.

[Lin *et al.*, 2003] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, 2003.

[Liu *et al.*, 2017] Fagui Liu, Dacheng Deng, and Ping Li. Dynamic context-aware event recognition based on markov logic networks. *Sensors*, 17(3):491, 2017.

[Loke, 2006] Seng Wai Loke. On representing situations for context-aware pervasive computing: six ways to tell if you are in a meeting. In *Fourth Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOMW'06)*, pages 5–pp. IEEE, 2006.

[Loke, 2016] Seng W Loke. Representing and reasoning with the internet of things: a modular rule-based model for ensembles of context-aware smart things. *EAI endorsed transactions on context-aware systems and applications*, 3(8), 2016.

[McCarthy and Hayes, 1981] John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.

[Mueller, 2004] Erik T Mueller. Event calculus reasoning through satisfiability. *Journal of Logic and Computation*, 14(5):703–730, 2004.

[Mueller, 2008] Erik T Mueller. Event calculus. *Foundations of Artificial Intelligence*, 3:671–708, 2008.

[Mueller, 2014] Erik T Mueller. *Commonsense reasoning: an event calculus based approach*. Morgan Kaufmann, 2014.

[Nelson and Mateas, 2008] Mark J Nelson and Michael Mateas. Recombinable game mechanics for automated design support. In *AIIDE*, 2008.

[Papadopoulos and Tzanetakis, 2016] Helene Papadopoulos and George Tzanetakis. Models for music analysis from a markov logic networks perspective. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):19–34, 2016.

[Patel *et al.*, 2002] Pranav Patel, Eamonn Keogh, Jessica Lin, and Stefano Lonardi. Mining motifs in massive time series databases. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 370–377. IEEE, 2002.

[Perera *et al.*, 2013] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. Context aware computing for the internet of things: A survey. *IEEE communications surveys & tutorials*, 16(1):414–454, 2013.

[Poon and Domingos, 2006] Hoifung Poon and Pedro Domingos. Sound and efficient inference with probabilistic and deterministic dependencies. In *AAAI*, volume 6, pages 458–463, 2006.

[Richardson and Domingos, 2006] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

[Russel *et al.*, 2013] Stuart Russel, Peter Norvig, et al. *Artificial intelligence: a modern approach*. Pearson Education Limited, 2013.

[Schmidt, 2003] Albrecht Schmidt. *Ubiquitous computing-computing in context*. Lancaster University (United Kingdom), 2003.

[Shanahan and others, 1997] Murray Shanahan et al. *Solving the frame problem: a mathematical investigation of the common sense law of inertia*. MIT press, 1997.

[Shanahan, 2004] Murray Shanahan. An attempt to formalise a non-trivial benchmark problem in common sense reasoning. *Artificial intelligence*, 153(1-2):141–165, 2004.

[Shanahan, 2006] Murray Shanahan. Frame problem, the. *Encyclopedia of cognitive science*, 2006.

[Shoham and McDermott, 1988] Yoav Shoham and Drew McDermott. Problems in formal temporal reasoning. *Artificial Intelligence*, 36(1):49–61, 1988.

[Singla and Domingos, 2005] Parag Singla and Pedro Domingos. Discriminative training of markov logic networks. In *AAAI*, volume 5, pages 868–873, 2005.

[Skarlatidis and Michelioudakis, 2014] Anastasios Skarlatidis and Evangelos Michelioudakis. Logical Markov Random Fields (LoMRF): an open-source implementation of Markov Logic Networks, 2014.

[Skarlatidis *et al.*, 2015] Anastasios Skarlatidis, Georgios Paliouras, Alexander Artikis, and George A Vouros. Probabilistic event calculus for event recognition. *ACM Transactions on Computational Logic (TOCL)*, 16(2):1–37, 2015.

[Snidaro *et al.*, 2015] Lauro Snidaro, Ingrid Visentini, and Karna Bryan. Fusing uncertain knowledge and evidence for maritime situational awareness via markov logic networks. *Information Fusion*, 21:159–172, 2015.

[Sutton and McCallum, 2006] Charles Sutton and Andrew McCallum. An introduction to conditional random fields for relational learning. *Introduction to statistical relational learning*, 2:93–128, 2006.

[Yeh *et al.*, 2016] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1317–1322. Ieee, 2016.

[Zhang *et al.*, 2019] Pei Zhang, Shijia Pan, Mostafa Mirshekari, Jonathon Fagert, and Hae Young Noh. Structures as sensors: Indirect sensing for inferring users and environments. *Computer*, 52(10):84–88, 2019.

Bartle Taxonomy-based Game for Affective and Personality Computing Research

Laura Żuchowska[1] , **Krzysztof Kutt**[2*] and **Grzegorz J. Nalepa**[1,2]

[1]Department of Applied Computer Science, AGH University of Science and Technology, Kraków, Poland
[2]Jagiellonian Human-Centered Artificial Intelligence Laboratory (JAHCAI) and Institute of Applied Computer Science, Jagiellonian University, Kraków, Poland
krzysztof.kutt@uj.edu.pl, gjn@gjn.re

Abstract

The paper presents the design of a game that will serve as a research environment in the BIRAFFE series experiment planned for autumn 2021, which uses affective and personality computing methods to develop methods for interacting with intelligent assistants. A key aspect is grounding the game design on the taxonomy of player types designed by Bartle. This will allow for an investigation of hypotheses concerning the characteristics of particular types of players or their stability in response to emotionally-charged stimuli occurring during the game.

1 Introduction and Motivation

Affective Gaming (AfG) [Lara-Cabrera and Camacho, 2019] is an area of research concerned with how games can measure and detect player emotions, and then use this information to adapt the game environment accordingly. If these modifications are also aimed at directing the player's affective state, e.g., towards specific emotions desired at a given stage of the game, then one can call this an affective feedback loop in which the game and the player interact (see Fig. 1).

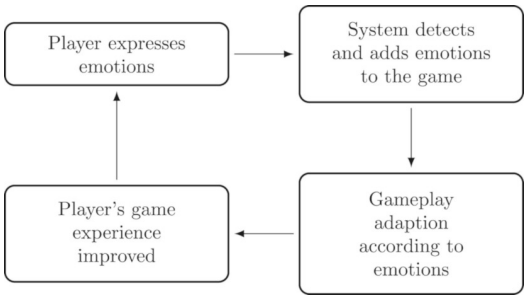


Figure 1: Affective game feedback loop [Lara-Cabrera and Camacho, 2019].

Studies in the AfG area do not just focus on entertainment. It can also be part of research projects concerning education [Dormann *et al.*, 2013] or the design of intelligent assistants, as in the BIRAFFE series of experiments [Kutt *et al.*, 2021a]. In the latter case, games are used as a fully controllable experimental environment that allows for accurate monitoring of the user's interaction with the system [Żuchowska *et al.*, 2020]. This is possible due to the similarities in human-in-the-loop [Nunes *et al.*, 2015] and affective loop interaction schemes. However, in order to extend the results of AfG research to interaction models of intelligent assistants in the future, careful game design and a system for logging the whole game context are required [Kutt *et al.*, 2021b].

The notion of context is understood as a component of emotion, according to the theory proposed by Prinz [2006]. In this view, context is anything that allows one to interpret a particular physiological activation and give it an appropriate interpretation. In the BIRAFFE series of experiments, the primary contextual information is behavioral data describing the interaction with the system/game – both the user's actions and the stimuli appearing in the system/game. In addition, demographic information (gender, age) and personality profiles are collected. Ultimately—when we move from a game-based experimental environment to real-world intelligent assistants—external sources of context, e.g., calendar data, current weather, will also be used. Importantly, once we have refined the low-level context storage mechanisms described in this paper, we also plan to attempt to derive higher-level context from them, e.g., instead of relying on changes in the position of individual characters in the game, we will operate on the information "the player is attacking an enemy" or "the player is running away from an enemy" instead.

This paper summarises the work carried out to prepare the game for the third experiment in the BIRAFFE (*Bio-Reactions and Faces for Emotion-based Personalization*) series. The motivation for developing the game in question was twofold. The first intention was to improve the experimental environment based on lessons learned from previous studies [Kutt *et al.*, 2021a; Kutt *et al.*, 2021b], in particular to provide a more accurate game context logging system. The second motivation was to extend the game design to include different types of interaction for different types of players. Combining information on the said types with personality profiles and physiological characteristics—obtained in all BIRAFFE experiments—will enable broader analyses that could lead to the identification of a set of characteristics for each type of player. It will also allow to investigate if and how the types and characteristics of users change during the course of the

*Corresponding Author

game.

The rest of the paper is organized as follows. In Sect. 2, Bartle taxonomy of player types is introduced. The design of the game with multiple paths for all player types is discussed in Sect. 3. Then, in Sect. 4 the set of logged contextual information is described. The paper is concluded in Sect. 5.

2 Bartle Taxonomy

The Bartle taxonomy [Bartle, 1997] is created on a 2D space, where the X axis is described as "Player – World", meaning the involvement of real people instead of non-playable characters (NPC) or world exploring in any way possible. The Y axis is set as "Acting – Interacting", which directly implicates the preference for acting or interacting. Each quarter of the space defines a different type of player as presented on Fig. 2.

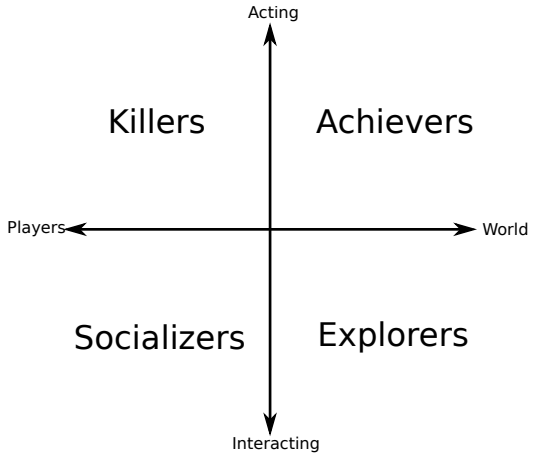


Figure 2: Bartle's taxonomy of player types [source: https://en.wikipedia.org/wiki/Bartle_taxonomy_of_player_types].

The achievers' goal is to *act* within the *world*. They wish to master the game, find the best possible weapon, get all the points (or achievements). People who take care about ranking and hierarchy can be considered achievers, therefore every competitive player is most likely an achiever. Another famous concept for achievers playing type is *grinding* – playing a game as long as it requires to get a desired outcome [Hilgard *et al.*, 2013].

Explorers start at simple exploring a topology of a game (breadth) and end at breaking the laws of in-game physics (depth), searching and using bugs. They are interested in *interacting* with the *world*. This player type is searching for knowledge and likes to be praised by others for having it. While game glitches are fun, players who like to find a specific, unique places and interesting features are also considered explorers. Additionally, speed-runners can also be labeled as a mix between achievers (if ranking is involved) and explorers.

For socializers, the most important part of the game is community and people, relations with them and interactions. They love to talk, sympathize and joke with others and appreciate the significance of *interacting* with *players*. For some socializers, observing the gameplay is enough. For others, some minor exploration can be included, in order to understand what other person is referring to. The act of killing is not required nor wanted for such a person to have fun. For single-player games, the socializer can be more entertained by making interesting NPC's with interesting backstory, multiple dialogue options and arcs, thought-provoking and engaging plot can be enough for socializer too.

Killers are a very specific, narrow group of people. Intentions behind a mind of a killer are clear to some extent. They enjoy being superior and high in hierarchy, however this is not in the same nature as achievers. Killers tend to do things they wouldn't normally do in real life, varying from punching a person to brutal murder. They also cherish the fact that they can do something to real human, who feels emotions and reacts, instead of NPC. Enjoyment comes from *acting* on *people*. Killers see other people, especially achievers, who can face the challenge, as their prey.

3 Game Design with Multiple Paths for Bartle's Player Types

The main goal of the game is to use the knowledge of player types in order to get closer to creation of a truly affective experience. As the BIRAFFE3 experiment is aimed to check the associations between the gamer's personality traits, physiological characteristics and in-game decisions (as introduced in Sect. 1), the proposed game provides an open world with as much non-linearity as possible [Gary, 2018]. The affectivity of the game has also been taken into account in the design – important choices will be accompanied by emotionally evocative stimuli, both sounds and images.

This game is fairly different than previous ones [Kutt *et al.*, 2021a; Żuchowska *et al.*, 2020], as it provides a pleasant gaming experience – something for everyone, no matter if a skilled player or casual person with no gaming background. Multiple point-increasing interaction systems have been introduced, such as dialogues with in-game characters (see Fig. 3). The story of each character is very simple, but rewarding enough to keep it entertaining for a subject [Torta and Minuty, 2017]. Some tasks and quests can be done for NPCs, mostly in a cute-bubbly way. In order to achieve that, all interactable, pickable objects have a type – consumable, plot, weapon or non-consumable. The next important interaction type is attack, which allows to kill an NPC or an animal in game with a previously found and equipped weapon item. It is important to notice that there is no difference in points added, whether the action is peaceful or not, the outcome in terms of points is always the same.

The sole purpose of the aforementioned affective pictures and sounds is to induce certain emotions in players, and see their reactions – the images and sounds will be displayed after some actions have been made. One of the most important activities, resulting in revealing a questionable image and/or sound, is chest opening. Opening such a special chest is one of many ways to gather points, however there is a trick to it. There are three types of chests: one with pleasant sounds and images, second one with 50:50 ratio to get a pleasant or disgusting image, and the third one which always displays an unpleasant, gore image. Every chest varies in terms of

Figure 3: Dialogue with the NPC.

Figure 4: Information about completed achievement.

amount of points it gives, which may result in an interesting insight on the subject's importance of points and horrible image watching. Of course, some people might not be interested in gathering points in the first place, which creates a challenge to overcome, as the images are a crucial part of affective experience. As far as Bartle taxonomy is considered, all types of gamers will find a way to see an affective image and hear a sound on a regular basis during the gameplay. Another ways to get the subject to look at such a picture include displaying an UI interface by talking with non-playable characters or reading boards and interacting with objects. After some random number of lines of text has been displayed, an image will be displayed in the background, however there will be no points for that, and the image will be random. Additionally, when achievement is unlocked by the player, depending on its type, a pleasant or undesirable sound will be played.

The whole game design was made specifically with a view to pursue the characteristics of each player type from Bartle taxonomy. Achievers can find multiple weapons and gather points, look up into current statistics and collect achievements for certain actions. The amount of points gathered thorough the game is being shown all the time in top left corner of the screen. Achievements on the other hand, are only displayed with the moment of completion (see Fig. 4). The first achievement will be very simple, in order to show that achievement gathering is possible, triggering some emotions in subjects with particular tendencies. Explorers will be interested by searching for hidden objects on the map and exploiting the mechanics, as some places have intentionally placed "bugs" as easter eggs. One of those bugs is an askew collider for map. In the bottom left corner of the game, there is a possibility to get out of the map (see Fig. 5) and find a hidden board with a nice message written on it (see Fig. 6). As for socializers, NPC are introduced, with their own backstories and problems to solve. Action with an NPC triggers an UI with dialogue options (see Fig. 3), allowing to know the character better and have a conversation. Killers can find pleasure in killing everybody around and committing acts that would be considered illegal or immoral in real life.

Figure 5: Intentionally placed "bug" in collider, allowing to get out of the map.

Figure 6: Hidden board placed out of the map as an "easter egg".

Technically, according to the assumptions made, the gameplay time should last 15 minutes. After that time, the game will end and proceed with the experimental procedure (as in other BIRAFFE experiments, see, e.g., [Kutt *et al.*, 2021a]). There is no possibility to finish game earlier, however there is nothing that keeps the subject from just standing in place for 15 minutes and stare blankly at the screen. The whole game was developed with the Unity Engine (https://unity.com/).

4 Logging System

To conduct a study based on such game, a suitable log handling had to be added. Similarly to the previous research [Kutt *et al.*, 2020], logs are created for each subject, based on their ID defined at the beginning of the experiment. A proper directory is created, along with all files about the game. During the gameplay, data containing current state of the player an the progress is being gathered with 10 Hz frequency. A log with an ID of subject as the name is written into JSON file and is being saved in application persistent data path. Such a log consists of various information about current state of the game:

1. Timestamp,
2. Location – both X and Y coordinates and area,
3. List of unlocked achievements,
4. Amount of interaction button clicks,

Figure 7: Trigger colliders for area logging information.

5. Number of interactions with unique objects,
6. List of particular milestones for NPCs tasks and dialogues,
7. If talking – name of the NPC, else an empty string,
8. Amount of killed NPCs,
9. Current equipped weapon,
10. Points and health,
11. List of items gathered,
12. List of opened chests,
13. ID of played sound and image.

The log file can be separated into groups. The first two items (items 1-2 on the list) are purely about the position over time of the protagonist, which may help with visualization or classification of commonly walked places in game. The "area" is a term describing important places in the world identified by arbitrarily prepared colliders (see Fig. 7). Second group (items 3-12) contains the characteristics of players behavior – did the protagonist gather achievements? Was s/he talking with NPCs? Maybe the subject was killing them? If so, with which weapon? How many points were gathered, etc. This section of logging system is supposed to help in analysis the most, as the heart of information about a pattern of playing. The last item (13) is for affect-related analyses – the ID of sound and image displayed after event.

Another log file contains the data about current state of the world. The characters are moving all the time, therefore their location needs to be written down as well – the position of each character can have an impact on each gameplay.

Finally, the last file, which is the same for all players, is the static map of the game world. It consists of information about the starting position of items, colliders, houses, etc. It's purpose is to allow for possible future visualization of events and analysis of collider interactions between the player and the world.

Keeping the Bartle taxonomy in mind, the log can be also separated into items related to specific gamer types. In terms of achievers, the information about points gathered and achievements unlocked is written, along with particular milestones for NPC's quests. The latter can also be used as a socializer trait, which is why the data on dialogue options clicked is also being saved – who was the player talking to. As for the explorers, the amount of unique objects interacted with together with the amount of interaction button clicks, chests opened and list of items is written into the file. Finally, for the killers, the data on the amount of NPC killed and type of equipped weapon is logged.

5 Summary and Future Work

The BIRAFFE series of experiments, which has been running for several years, focuses on the development of interaction models for personalised intelligent assistants based on a range of contextual information about the user: physiological signals collected with low-cost wearable devices, personality assessment, behavioural data describing the interaction with the system, and external sources of context (such as current weather conditions). A means to the goal is to use games as a stimulus-rich yet fully controllable experimental environment.

This paper presents the design of a new affective game to be used in the BIRAFFE3 experiment, scheduled for autumn 2021. In addition to addressing the weaknesses found in previous games, a new contribution of using Bartle's taxonomy during interaction design is introduced. This will enable post-experimental analyses focusing on determining the characteristics of specific user types or investigating the stability/variability of player type in response to positive/negative stimuli associated with their in-game interactions. We believe that inclusion of Bartle player types into both the design of the affective game, as well as data analysis about player interaction with it, provides a new and important source of context.

Finally, the post-experimental analyses will also focus on creating a catalogue of interaction patterns, which will be the basis for creating an improved version of the game, allowing the gameplay to adapt to the player's emotions, i.e., implementing a full affective game feedback loop. This will thus allow a transition from a "Detection and measure" approach to an "Integral approach" according to the Lara-Cabrera and Camacho's taxonomy [2019].

Acknowledgements

The research has been supported by a grant from the Priority Research Area Digiworld under the Strategic Programme Excellence Initiative at the Jagiellonian University.

References

[Bartle, 1997] Richard Bartle. Hearts, clubs, diamonds, spades: Players who suit muds. *The Journal of Virtual Environments*, 1(1), 1997.

[Dormann *et al.*, 2013] Claire Dormann, Jennifer R Whitson, and Max Neuvians. Once more with feeling: Game design patterns for learning in the affective domain. *Games and Culture*, 8(4):215–237, 2013.

[Gary, 2018] Justin Gary. *Think Like a Game Designer*. Aviva Publishing, Lake Placid, NY, 2018.

[Hilgard *et al.*, 2013] Joseph Hilgard, Christopher Engelhardt, and Bruce Bartholow. Individual differences in motives, preferences, and pathology in video games: the gaming attitudes, motives, and experiences scales (GAMES). *Frontiers in Psychology*, 4:608, 2013.

[Kutt *et al.*, 2020] Krzysztof Kutt, Dominika Drążyk, Maciej Szelążek, Szymon Bobek, and Grzegorz J. Nalepa. The BIRAFFE2 experiment. study in bio-reactions and faces for emotion-based personalization for AI systems. *CoRR*, abs/2007.15048, 2020.

[Kutt *et al.*, 2021a] Krzysztof Kutt, Dominika Drążyk, Szymon Bobek, and Grzegorz J. Nalepa. Personality-based affective adaptation methods for intelligent systems. *Sensors*, 21(1):163, 2021.

[Kutt *et al.*, 2021b] Krzysztof Kutt, Laura Żuchowska, Szymon Bobek, and Grzegorz J. Nalepa. People in the context – an analysis of game-based experimental protocol. In *MRC@IJCAI 2021*, 2021. in press.

[Lara-Cabrera and Camacho, 2019] Raúl Lara-Cabrera and David Camacho. A taxonomy and state of the art revision on affective games. *Future Generation Computer Systems*, 92:516–525, 2019.

[Nunes *et al.*, 2015] David Sousa Sousa Nunes, Pei Zhang, and Jorge Sá Silva. A survey on human-in-the-loop applications towards an internet of all. *IEEE Commun. Surv. Tutorials*, 17(2):944–965, 2015.

[Prinz, 2006] Jesse J. Prinz. *Gut Reactions. A Perceptual Theory of Emotion*. Oxford University Press, Oxford, 2006.

[Torta and Minuty, 2017] Stephanie Torta and Vladimir Minuty. *Storyboarding: Turning Script into Motion*. Mercury Learning and Information, Dulles, VA, 2017.

[Żuchowska *et al.*, 2020] Laura Żuchowska, Krzysztof Kutt, Krzysztof Geleta, Szymon Bobek, and Grzegorz J. Nalepa. Affective games provide controlable context. proposal of an experimental framework. In Jörg Cassens, Rebekah Wegener, and Anders Kofod-Petersen, editors, *Proceedings of the Eleventh International Workshop Modelling and Reasoning in Context co-located with the 24th European Conference on Artificial Intelligence, MRC@ECAI 2020, Santiago de Compostela, Galicia, Spain, August 29, 2020*, volume 2787 of *CEUR Workshop Proceedings*, pages 45–50. CEUR-WS.org, 2020.