Tutorial: developing a data annotation protocol

Teodor Stoev
Institute for Data Science, University of Greifswald
Greifswald, Germany
teodor.stoev@uni-greifswald.de

Emma L. Tonkin
University of Bristol
Bristol, UK
e.l.tonkin@bristol.ac.uk

Kristina Yordanova
Institute for Data Science, University of Greifswald
Greifswald, Germany
kristina.yordanova@uni-greifswald.de

Gregory J. L. Tourte
University of Bristol
Bristol, UK
g.j.l.tourte@bristol.ac.uk

ABSTRACT

Data annotation is key to a large number of fields, including ubiquitous computing. Documenting the quality and extent of annotation is increasingly recognised as an important aspect of understanding the validity, biases and limitations of systems built using this data: hence, it is also relevant to regulatory and compliance needs and outcomes. However, the process of annotation often receives little attention, and is characterised in the literature as 'under-described' and 'invisible work'. In this tutorial, we bring together existing resources and methods to present a framework for the iterative development and evaluation of an annotation protocol, from requirements gathering, setting scope, development, documentation, piloting and evaluation, through to scaling-up annotation processes for a production annotation process. We also explore the potential of semi-supervised approaches and state-of-the-art methods such as the use of generative AI in supporting annotation workflows, and how such approaches are validated and their strengths and weaknesses characterised. This tutorial is designed to be suitable for people from a wide range of backgrounds, as annotation can be understood as a highly interdisciplinary task and often requires collaboration with subject matter experts from relevant fields. Participants will trial and evaluate a selection of annotation interfaces and walk through the process of evaluating the outcomes. By the end of the workshop, participants will develop a deeper understanding of the task of developing an annotation protocol and aspects of the requirements and context which should be taken into account.

Presentations and code from this event will be shared openly on a Github repository.

KEYWORDS

data annotation, data labelling, manual annotation, automated annotation

INTENDED AUDIENCE

Data annotation is broadly useful across a wide variety of domains, and as such, this tutorial is designed to be accessible to participants from a wide variety of backgrounds, from engineering and machine learning to design and human-computer interaction. We aim to demonstrate representative examples of tools and approaches to evaluation within a framework of best practices, with the aim of providing decision-makers and potential implementors of data annotation with a toolkit that facilitates practical decision-making in this area. Some familiarity with Python or R may be useful for the evaluation of annotation practices. However, we will provide accessible examples.

The aim of this tutorial is to provide a step-by-step framework through which to approach data annotation tasks, from initial conceptualisation and piloting through to evaluation of results. As such, we encourage participants with existing plans for annotation tasks to bring a summary of their planned project, as there will be an opportunity to work through these in groups during the workshop.

1 MOTIVATION

Data annotation plays a crucial role in the development and evaluation of modern AI systems. Especially in domains such as activity recognition or situation-aware user assistance where wearable and non-wearable sensors are used to collect data, it is important to have reliably labelled samples in order to perform training and validation of the underlying models. However, data annotation is not only about training and assessing algorithms or tagging a set of data samples, but an attempt to meet the requirements of multiple stakeholders, such as system engineers, machine learners, people making use of the system and people tasked with elements of the annotation process. Hence, data annotation, the encoding of information about the world to support the use of data, is an interdisciplinary engineering task.

The quality of annotation is often an important defining factor in the success of engineering processes. Annotation is often time-consuming and expensive [9] and hence a great deal of research has focused on how to limit these costs, whether through improved interface design, technical measures such as preprocessing and automation [2] or through outsourcing, such as crowdsourcing models (for example, the MODA platform [3] or crowdsourcing marketplaces [4]) or outsourcing to third-party data labelling firms [8]. As Wang et al. [8] notes, data annotation labour is under-recognised in general,and this 'invisible' work at times leads to exploitative practices and harms. Annotation, then, is an under-discussed and significant component of many engineering tasks. It is also increasingly likely to have regulatory impact, as nations have begun to look at mitigations of risks associated with machine learning applications: for example, the draft EU AI Act [1] lays down requirements for training, validation and test datasets which 'should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system. They should also have the appropriate statistical properties, including as regards the persons or groups of persons on which the high-risk AI system is intended to be used. In particular, training, validation and testing

data sets should take into account, to the extent required in the light of their intended purpose, the features, characteristics or elements that are particular to the specific geographical, behavioural or functional setting or context within which the AI system is intended to be used.'

2 OUTLINE

This tutorial aims to motivate participants to delve into the world of data annotation and acquire a broad understanding of the fundamental components involved in the annotation process. By examining a series of worked examples within a cohesive framework, attendees will acquire knowledge in the following topics: what is data annotation and how to create annotation guidelines, how to choose the right tools for the particular data at hand, how to estimate the costs of annotation and which effective annotation techniques exist. In particular, participants will gain experience in developing a worked plan for data annotation, starting with the elicitation of initial requirements and working through testing, piloting and evaluation phases, making use of opportunities for input from the various stakeholders involved and taking into account the practical constraints of the project and the perspectives of the data annotators themselves. Participants will work in groups to explore scenarios of use and share their insights. We invite people from different backgrounds and fields to join the tutorial. Attendees are not expected to have experience in the field of data annotation.

In the tutorial we will introduce the topic of data annotation and some of its main uses in the context of ubiquitous computing. By working through real-world examples in groups, including those proposed by participants, we will discuss requirements gathering for annotation, defining the scope of annotation tasks, evaluation of candidate approaches and the development, evaluation and iteration of annotation guidelines and processes. We will explore some of the many tools available to facilitate specific workflows or types of data annotation, and the role of effective requirements analysis in the design and evaluation of annotation approaches and in development of an effective annotation protocol. Finally, we will discuss means to increase the efficiency of data annotation tasks and alternatives to intensive data annotation, as well as the role of generative AI, such as Large Language Models (LLMs).

For the tutorial we will prepare a dedicated github repository with all the materials needed for the practical sessions.

2.1 Topics covered

- Requirements gathering
  - understanding your data
  - understanding what you are expecting to extract out of the data, and the purposes for which it is collected
  - knowledge elicitation – annotation type, domain and coverage
  - exploring available annotation tools
  - annotation modalities – participant self-annotation, contemporaneous annotation, post-hoc annotation
  - format (just as the actual data, the formats for the annotation is important and will vary depanding on the workflow). Take into account how the annotations will be used (manual, programs), if will be shared/ published
  - who is annotating and when (will depend on the types and doesn't have to be the same for each type of data within the same project)
  - understanding data protection implications
  - understanding sensitivity of the tasks for annotators and potential for harm
  - risk assessment process highlighting ethical consideration of annotation process
  - understanding regulatory implications (documenting bias, limitations and risks of harm)
- Scope definition
  - granularity (what you need to describe, what you could describe, and what you really don't need).
  - the role of exploratory annotations – conceptualising and formalising the annotation task through exploratory note-taking or interaction with data [5]
  - limitations (no annotation is prefect, difference in interpretation of either the data or the annotation vocabulary, controlled or not; who is doing the annotation, self (subject of the data) or post hoc annotators).
- Method development and documentation
  - initial development of annotation guidelines [5] and a data annotation protocol
  - iterative improvement of guidelines as a result of piloting, testing and evaluation
- Testing and evaluation process
  - defining evaluation process and suitable annotation metrics
  - critically reviewing the solutions proposed from the annotation processes and scope definition steps, above
  - ensure the feasibility of the annotation solutions, understand the practical constraints (time, communication, required skillsets and background knowledge, costs)
  - evaluate the data labels – are they usable and suitable for the intended purposes of use? Testing against competency questions or sample tasks
  - evaluate inter – annotator consistency – Fleiss' and Cohen's kappas, and handling inconsistent numbers of annotators
  - evaluate the results in situ – e.g. does it fulfil the requirements within the intended context of use (intended data output)?
  - refine/change solution if needed, returning to previous steps (iterative process)
- Implementation
  - scaling up beyond the pilot
- Current research
  - how can we do more with less? Embedding methods such as active learning, weak supervision, soft labelling and recording of annotator attention and confidence.
  - the role of generative AI, such as LLMs, in annotation tasks – strengths, weaknesses and principled validation and use.

3 FORMAT AND SCHEDULE

We propose the following structure for our half-day tutorial:

- Introduction: This section will provide an overview of data annotation, its importance, and relevance in the AI industry.
- Types of data annotation: This section will explore the most common types of data annotation, including image annotation, text annotation, audio annotation and annotation of sensor data
- Developing annotation guidelines: The participants will learn how to design annotation guidelines and what strategies play an important role regarding speed and cost of annotation.
- Annotation quality: The participants will walk through examples of how to measure the quality of annotations and explore the role of visual analysis in understanding inconsistencies.
- The human factor in data annotation: We will discuss certain human factors which might affect the annotation process such as the emotional state of annotators, background, data sensitivity, time constraints, etc. We will also talk about how ethical considerations should be addressed during annotation.
- Annotation tools: We will discuss popular open source annotation tools, their features, and their suitability for different types of annotation data.
- Effective annotation techniques: We will delve into different annotation techniques such as active learning, crowdsourced annotation, and semi-supervised learning. It will also provide insights on how to implement these techniques effectively.
- Hands-on annotation: within groups, participants will be encouraged to annotate a small amount of different types of data with open source tools of their choice, and to discuss and document the pros and cons of using particular tools as well as identifying possible future improvements.
- Conclusion: At the end of our tutorial we will provide a summary of the tutorial and highlight the importance of data annotation in the AI development process.

By the end of this tutorial, participants will have a solid understanding of data annotation, its types, tools, and effective annotation techniques. They will be able to apply their knowledge in the AI development process, making them a valuable asset in industry and/or research.

4 SUPPORT MATERIALS

Participants are encouraged to bring a laptop to the tutorial. The tutorial materials will be placed online, in a git repository, for future reference, with accompanying code and example templates.

5 ORGANISERS' BACKGROUNDS

The organisers of this tutorial have a broad expertise in the domain of data annotation. All of them have been involved in the organization of a series of full-day workshop events called ARDUOUS (Annotation of useR Data for UbiquitOUs Systems), affiliated to the IEEE International Conference on Pervasive Computing and Communications (PerCom)). ARDUOUS was put together to support the development of a community of practitioners working in annotation for pervasive systems research and over the period of seven years (2017–2023 [11, 10, 12, 7, 14, 13]) a broad spectrum of topics related to data annotation has been discussed. A short biography of each organiser is included below.

Dr Emma L. Tonkin is a research fellow at the University of Bristol, UK. She is an expert in Digital Humanities and Data Science for Digital Health. She is part of the SPHERE project (Sensor Platform for HealthcarE in a Residential Environment), where she works on data management and analysis, data reproducibility and data ethics. Within the Mare Balticum Fellowship Programme[1] of the University of Rostock, Dr Emma L. Tonkin was awarded a visiting fellowship during which she gave a series of lectures on data management, privacy, and annotation.

Prof. Dr.-Ing. Kristina Yordanova is the head of the Institute for Data Science at the University of Greifswald, Germany. Previously she was the group leader of the junior research group CoMSA²t (Cognitive Methods for Situation-Aware Assistive Systems). She is also a research associate at the University of Bristol in the context of the SPHERE project. Her research interests are in the fields of activity and intention recognition, human behaviour models, knowledge elicitation, natural language processing, and automatic extraction of behaviour models from textual sources. She is one of the initiators of the ARDUOUS workshop series and has many years of experience with data annotation.

Teodor Stoev is a research assistant and a PhD student who is working in the project BeHave[2] [6]. His research interests include knowledge-based models, domain knowledge extraction from heterogeneous data sources, sensor data analysis, machine learning, artificial intelligence, and data annotations. He co-chaired the ARDUOUS2021 and ARDUOUS2022 workshops.

Gregory Tourte is Senior Research Associate at the University of Bristol in the School of Geographical Sciences where he started as a research software engineer and system administrator for the research group's supercomputer and used this opportunity to develop an understanding of research data management due to the large quantity of data being generated. He continues his work with deep time climate modelling within the Bristol Research Initiative for the Dynamic Global Environment (BRIDGE). In March 2023 he co-chaired the 7th ARDUOUS workshop affiliated to PerCom 2023.

REFERENCES

[1] Artificial Intelligence Act. 2021. Proposal for a regulation of the european parliament and the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *EUR-Lex-52021PC0206*.

[2] Patrícia Bota, Joana Silva, Duarte Folgado and Hugo Gamboa. 2019. A semi-automatic annotation approach for human activity recognition. *Sensors*, 19, 3, 501.

[3] Karine Lacourse, Ben Yetton, Sara Mednick and Simon C. Warby. 2020. Massive online data annotation, crowdsourcing to generate high quality sleep spindle annotations from eeg data. *Scientific Data*, 7, 1, (19th June 2020), 190. DOI: 10.1038/s41597-020-0533-4.

[1] https://www.inf.uni-rostock.de/wkt/nachwuchsfoerderung/mare-balticum-fellowship-programm/

[2] https://www.mmis.informatik.uni-rostock.de/research/projects/behave/

[4] Tanushree Mitra, C. J. Hutto and Eric Gilbert. 2015. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). Association for Computing Machinery, Seoul, Republic of Korea, 1345–1354. ISBN: 9781450331456. DOI: 10.1145/2702123.2702553.

[5] Janis Pagel, Nils Reiter, Ina Rösiger and Sarah Schulz. 2018. A unified text annotation workflow for diverse goals. In *Sandra Kübler/Heike Zinsmeister (Hg.), Proceedings of the Workshop on Annotation in Digital Humanities, co-located with ESSLLI*, 31–36.

[6] Teodor Stoev and Kristina Yordanova. 2021. Behave: behaviour understanding through automated generation of situation models. In *KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings*. Springer, 362–369.

[7] Emma L. Tonkin and Kristina Yordanova. 2020. Arduous 2020: 4th international workshop on annotation of user data for ubiquitous systems – welcome and committees. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE Computer Society, Los Alamitos, CA, USA, (Mar. 2020), 1–2. DOI: 10.1109/PerComWorkshops48775.2020.9156077.

[8] Ding Wang, Shantanu Prabhat and Nithya Sambasivan. 2022. Whose ai dream? in search of the aspiration in data annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22) Article 582. Association for Computing Machinery, New Orleans, LA, USA, 16 pages. ISBN: 9781450391573. DOI: 10.1145/3491102.3502121.

[9] Jingru Yang, Ju Fan, Zhewei Wei, Guoliang Li, Tongyu Liu and Xiaoyong Du. 2018. Cost-effective data annotation using game-based crowdsourcing. *Proceedings of the VLDB Endowment*, 12, 1, 57–70.

[10] Kristina Yordanova and Adeline Paiement. 2018. Percom workshops 2018 committees. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, i–ii. DOI: 10.1109/PERCOMW.2018.8480146.

[11] Kristina Yordanova, Adeline Paiement, Max Schröder, Emma L. Tonkin, Przemysław Woznowski, Carl Magnus Olsson, Joseph Rafferty and Timo Sztyler. 2018. Challenges in annotation of user data for ubiquitous systems: results from the 1st arduous workshop. *CoRR*, abs/1803.05843. arXiv: 1803.05843 [cs.CY]. DOI: 10.48550/arXiv.1803.05843.

[12] Kristina Yordanova, Emma L. Tonkin and Adeline Paiement. 2019. Arduous 2019 – 3rd international workshop on annotation of user data for ubiquitous systems – welcome and committees. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE Computer Society, Los Alamitos, CA, USA, (Mar. 2019), 1–2. DOI: 10.1109/PERCOMW.2019.8730683.

[13] Kristina Yordanova, Emma L. Tonkin and Teodor Stoev. 2022. Annotation of user data for ubiquitous systems – welcome and committees. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE Computer Society, Los Alamitos, CA, USA, (Mar. 2022), i–ii. DOI: 10.1109/PerComWorkshops53856.2022.9767338.

[14] Kristina Yordanova, Emma L. Tonkin and Teodor Stoev. 2021. Arduous 2021: 5th international workshop on annotation of user data for ubiquitous systems – welcome and committees. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 1–2. DOI: 10.1109/PerComWorkshops51409.2021.9430959.