philosophies

MDPI

Essay

Conversational Systems in Machine Learning from the Point of View of the Philosophy of Science—Using Alime Chat and Related Studies

Tosin P. Adewumi *, Foteini Liwicki and Marcus Liwicki

Department of Computer Science, Electrical and Space Engineering, EISLAB, Machine Learning Group, Luleå University of Technology, 97187 Luleå, Sweden

* Correspondence: oluwatosin.adewumi@ltu.se

Received: 25 June 2019; Accepted: 18 July 2019; Published: 23 July 2019

Abstract: This essay discusses current research efforts in conversational systems from the philosophy of science point of view and evaluates some conversational systems research activities from the standpoint of naturalism philosophical theory. Conversational systems or chatbots have advanced over the decades and now have become mainstream applications. They are software that users can communicate with, using natural language. Particular attention is given to the Alime Chat conversational system, already in industrial use, and the related research. The competitive nature of systems in production is a result of different researchers and developers trying to produce new conversational systems that can outperform previous or state-of-the-art systems. Different factors affect the quality of the conversational systems produced, and how one system is assessed as being better than another is a function of objectivity and of the relevant experimental results. This essay examines the research practices from, among others, Longino's view on objectivity and Popper's stand on falsification. Furthermore, the need for qualitative and large datasets is emphasized. This is in addition to the importance of the peer-review process in scientific publishing, as a means of developing, validating, or rejecting theories, claims, or methodologies in the research community. In conclusion, open data and open scientific discussion fora should become more prominent over the mere publication-focused trend.

Keywords: conversational systems; chatbots; philosophy of science; objectivity; verification; falsification

1. Introduction

In this essay, the authors discuss conversational systems (also called chatbots) of natural language processing (NLP) in machine learning (ML), from the philosophy of science point of view. The authors' position on the theory of how science operates is one of naturalism [1]. Hence, the objective of this essay is to evaluate conversational systems' research activities in light of this philosophical theory. This theory of knowledge is similar to the precept and example of, the now defunct, logical empiricism, which viewed only verifiable statements as meaningful [1]. Understanding of the way the world functions or the theory that explains observations may influence what is perceived. Just as the scientific community holds on to certain assumptions alluded to by Kuhn [2], the conversational systems community is not exempt from these assumptions. The assumptions, central to naturalism, are a collection of beliefs and values, untested by the scientific processes. They, however, give legitimacy to the scientific systems and set boundaries of investigations. One such basic assumption is that random sampling is representative for an entire population [3]. Possible benefits from this essay are that it summarizes improvements made in the science of developing conversational systems; and that it suggests that certain practices, such as the peer-review system and the use of qualitative, less biased, and large datasets, will bring further improvements.

Conversational systems are software systems that use natural language to communicate with users. This may be through written text or spoken dialogue [4]. The development of conversational systems began in the 1960s with Eliza being the product of such early studies [5]. This was a turning point in artificial intelligence (AI)—the imitation of human intelligence by software or hardware. AI is different from logical reasoning, problem-solving, or symbol's manipulation. However, some members of the AI community will agree that logic plays some role in the plethora of AI research areas [6]. Machine learning, which has become popular over the past few decades, is a subset of AI which is concerned with the learning of patterns for making predictions or performing specific tasks by using algorithms and statistical models without explicit programming [7]. The learning procedure takes place during training, with the aim of generalizing to 'unseen' data while avoiding overfitting (memorization) [8]. Natural language processing systems can be trained using text corpora (a large and structured set of texts) [9]. Examples of chatbots include Apple's Siri and Google Assistant.

Alime Chat is another chatbot developed by Alibaba researchers, mainly for Chinese [10]. It was developed for customer service operations at Alibaba[1] and can handle about 85% of the total customer service operation [10]. It is mainly a hybridized chatbot that leverages the capabilities of both information retrieval (IR) and machine learning generation models. Information retrieval and generation model approaches are categorized as data-driven because they rely mainly on data sources [11]. The latter synthesizes novel sentences, word by word, based on a dialogue history and persona (if included) [12,13]. Meanwhile, the information retrieval approach retrieves stored information, such as documents, images, speech, and video, from repositories [9,11,14]. The reason for selecting Alime Chat research and its related studies is because they mark new trends in conversational systems' problem solving and many of them are being used in industry as well. Indeed, Alime Chat currently answers millions of customers' questions per day at Alibaba.

When a philosophy of science outlook regarding a given research subject is taken, there are at least two possibilities: One being to look at the research activities in the discipline being studied and evaluate the various philosophical theories proposed about the functioning of science and its epistemological status; the second being to adhere to a particular theory of how science operates and choose to evaluate the discipline's activities against the chosen philosophical theory; we chose the second approach. In the following, you will find the methodological issues section, the exposition of the chosen studies section, and the summary and conclusion section. The methodological issues section summarizes the approach and some metrics used in conversational systems research, while the exposition section discusses some of the research activities from the point of view of the philosophy of science. Finally, the summary and conclusion section reiterates the main features of the discussion.

2. Methodological Issues

The methodology followed for gathering empirical data plays an important role. It must be unbiased (or impartial), as much as possible, and critical in its approach. Comparative studies, where performance of two or more systems are tested, are popular methods in conversational systems and they are usually based on experiments.

Various metrics (or measurements) exist in natural language processing. The BLEU score measures language translation success and was proposed by Papineni et al. [15]. It measures how closely machine translation is to standard human translation and how this correlates to human accuracy [15]. It is, however, reportedly not accurate when predicting single sentence human judgment, according to Lipton, Berkowitz, and Elkan [14] and therefore METEOR was introduced as an alternative. In addition, the GLEU score is an evaluation metric for sentence-level fluency [16].

[1] https://www.alibaba.com/

Human or manual evaluation is considered a better metric than any other, since human understanding of what is produced is what is ultimately sought [13]. Despite the benefits of this type of evaluation, it has disadvantages: it is costly and subjective [17,18]. It is costly in terms of resources (such as money and time) since the human subjects have to be recruited and trained before evaluation.

3. Exposition of the Chosen Studies

According to Thagard, when we can deduce statements, based on observation, from an occurrence, then a theory around such occurrence is verifiable [19]. For example, researchers in conversational systems, including Alime Chat, conduct several experiments and collect data by observation to make inferences [10,11]. Inference refers to the process of drawing conclusions, sometimes done after a statistical analysis is carried out. Statistical analysis is the evaluation of data for the purpose of inference [3]. There are three main types of inference: deduction, induction, and abduction. What is inferred is necessarily true in deductive inferences, given true premises. Meanwhile, the nature of induction and abduction is one of non-necessary inference [20]. In a comparative study method, two or more systems' performances are assessed based on certain defined metrics (such as BLEU or GLEU) and the better or worse system is established from the outcome of several observations, as an average. Hence, though it is possible in some observations to find cases where a system with a low performance performs better than a system with a high performance, this is not sufficient enough to question the preeminence of the better system. Such a case can merely be seen as an anomaly. This is because one or a few out of many cases is not enough to invalidate a position, since many instances were conducted to arrive at an average.

Methods of inquiry require objectivity in their approach. Objectivity, whose value and attainability has been repeatedly criticized in the philosophy of science, is usually regarded as the basis of the authority of science or the reason for valuing science [21]. It prescribes that the components of science (such as methods and claims) should not be influenced by personal interests, community bias, or other similar factors [21]. Product objectivity and process objectivity are the two basic ways of understanding objectivity. Product objectivity is based on science's theories, experimental results (e.g., BLEU scores), observations, and similar products constituting accurate representations of the world [21,22]. Process objectivity is multi-faceted and shows how science is objective to the point that the scientist's individual bias or contingent social values are not what science's processes and methods depend on [21]. An examination of the several conceptions of the ideal of objectivity is outside the scope of this essay. However, it has been argued that the facts of science are necessarily perspectival because of the involved apparatus and sociological factors [21]. Hence, given that full objectivity may not be deliverable, the conversational systems community plays a key role in describing what constitutes objectivity, which brings about trust in the science, as part of the social process. Indeed, Longino admitted that her analysis was not meant to be complete but to provide a starting framework from which the epistemologist (philosophers of the theory of knowledge) community could fill in further details [22].

Objectivity is a value which, as mentioned earlier, has been criticized extensively in the philosophy of science. Willingness to let the facts determine our beliefs, marks our objectivity. This is a position Longino does not seem to be averse to [22]. However, possible suspicion of what constitutes "the fact" from her submission, suggests that this needs to be carefully considered. For example, she suggests that the data used in a research experiment (which count as facts in that study) also need to be checked for reliability [22]. Hence, checking that the data has been interpreted by the authors in a subjective-free way is an important function in a peer-review process [22]. Furthermore, identification of possible institutional bias in the post-publication stage of a given idea was rightly identified by Longino [22]. This means that scientific publications should not be seen as the end. Attempts to reproduce experiments, subsequent use and modification by others are equally essential and can eventually compensate for institutional bias [22].

Conversational systems research makes use of the scientific method. The scientific method has process objectivity as its basis [21,22]. As Longino pointed out, the scientific method is the use of non-arbitrary and non-subjective criteria for developing, accepting, and rejecting a scientific view [22]. Since objectivity itself may not be fully attainable, this has an impact on scientific methods, and again, makes the role played by the conversational systems community relevant to prescribing what constitutes the scientific method. This view is supported by Longino, who identified two shifts in perspective related to the scientific method, the second shift being made possible by refocusing on "science as practice". In her work, she proposes that this involves the subjection of hypotheses and the background assumptions to varieties of conceptual criticism [22]. Her point about objectivity of scientific methods being a function of both observational data and background assumptions lends credence to practices in the conversational systems community [22]. Usually, the methods used in conducting experiments are provided for scrutiny, by researchers, to ensure their external and internal validity. Such information gives assurance to the conversational systems community about the objectivity of the results and the data used. Therefore, statistical analyses on such data can also be seen as objective. For example, Alime Chat researchers clearly stated the source of the data used, the architecture of the network, and the steps involved in producing the experiments [10]. This is also the case in a related study by Song et al. [11]. Furthermore, Longino observed that experiments based on unstable, quickly-evolving assumptions, lack objectivity [22]. Hence, observer effects, which may cause undue influence on research, are not objective. Methods employed in research should be a collection of social processes (such as the peer-review process for scientific publishing), as argued in [22]. This view is similar to Kuhn's position on the acceptance or rejection of a paradigm, which he argued should be a social process as much as a logical one [2].

In research on conversational systems, the type and size of data used for training influences the quality of the conversational systems created. For example, a small dataset utilized as an underlying corpus will produce poor performance when compared to a large dataset [9,11]. Similarly, a biased dataset (either being a stereotyped dataset or a partial dataset) will be reflected in the performance of a conversational system, as was witnessed with Microsoft's chatbot Tay, which posted racist comments and conspiracy theories online after having been exposed to data of users who (intentionally or unintentionally) exploited the chatbot's sensitivity by posting many racist comments and conspiracy theories [23]. After valuable discussion with the anonymous reviewers of this essay, we should add that it is, in general, difficult to create an unbiased dataset. Indeed, for machine learning, a bias is typically needed to actually learn something. The most crucial issue, however, is to remove unwanted/harmful biases, such as racist, gendered, societal discriminatory, or hate-speech entries. Furthermore, an example for creating a less biased dataset (in the context of an insurance company) would be taking all inquiries (not only made in chats, but also by phone calls and physical visits) made by all customers and randomly selecting a subset of that. Public fora, such as conferences, workshops, and journals, provide avenues for criticism of research and its constituent parts. It is also through such avenues that shared standards can be learned and responses to criticism given. Despite concerns (such as unwarranted blocking of publications) regarding the peer-review process in scientific publishing, it is considered a very useful system for evaluating the objectivity of research methods and claims made in scientific papers [22]. It is a useful filter system that assesses whether an article conforms to generally agreed guidelines provided by the research community. The various articles on conversational systems cited in this essay were published in peer-review journals, which means they had been subject to some critical evaluation or criticism by members of the scientific community before being published.

In refuting conjectures, Popper was opposed to the procedure of inference as a result of many observations [24]. However, usually, claims made in conversational systems research are based on evidence from observations. This approach raises the concern of how many observations are sufficient to avoid refutation, as expressed by Popper. Furthermore, Lipton categorically states that this approach cannot be taken as a proof of evidence [25]. Although abduction may be considered in a philosophical debate, the nature of the problem or debate plays an important part in its application, some even considering induction to be a special type of abduction [20]. Taking into account that we must be careful when concluding from empirical data, it is generally accepted that examples help in argument clarification and empirical confirmation and can increase the probability of the conclusion or claim. For example, Alime Chat researchers repeated 2136 tests in order to validate the obtained high performance of their system. Although Popper may have disagreed with this approach, the willingness of the conversational systems community to confirm or disconfirm their position, based on sufficient evidence, suggests that it is a reasonable approach. The willingness of the community to change, based on active research, is one of the scientific criteria alluded to by Thagard [19]. Lakatos may have approved this approach as the right one, since blind commitment is as serious a crime as any according to him [26]. Researchers in the area of conversational systems are not blindly committed to the claims or theories made, but are making strong efforts to ascertain the facts by reproducing experiments and are, in some cases, even advancing the field of research by trying out new methods. For instance, in determining if their hypothesis of a hybrid system was better, the Alime Chat researchers developed a new hybrid system and ran similar tests comparable to the old systems [10]. Song et al. similarly compared five architectures, including a baseline [11].

Confirmation by verification is not the only approach applicable in conversational systems, though this approach is sufficient for those who believe a theory is scientific only if it is verifiable [19]. The condition for refuting a claim can also be used. Popper states that in order for a claim or theory to be considered scientific, one should present a condition in which such a theory can be considered falsifiable or refutable [24,26]. Such a test can be applied to some of the claims made in the conversational systems research society. For example, in order to compare Alime Chat with another chatbot in production, the researchers conducted 878 experiments on each of the chatbots [10]. In order to falsify their claim that Alime Chat was better, the researchers argued that the other chatbot had to win by conversing better (when answering questions, as evaluated by humans) in a majority number of times.

4. Summary and Conclusions

Standards and processes for conducting research in the area of conversational systems have been improved through the plethora of avenues created by the research community. In this essay, it has been shown that the mentioned research area uses scientific methods in developing, accepting, and rejecting proposed theories using rational and non-subjective criteria, as posited by Longino [22]. Full objectivity may not be realizable because of the apparatus of science and sociological factors (such as biases); however, the conversational systems community plays a key role in describing which components constitute the objectivity that brings trust. Furthermore, the importance of confirmation by verification was mentioned, as well as the use of falsification, as stated by Popper [24].

The process of improving the methodology employed in conversational systems research is lively and continual. This is especially important because we must be cautious when drawing conclusions from empirical data. Empirical confirmation, however, increases the probability of claims. The need for qualitative data as well as large amounts of data was pointed out in this essay. It is difficult to completely eliminate bias from datasets; however, efforts should be made to eliminate the presence of unwanted bias or stereotypes, which can negatively influence the performance of conversational systems. In addition, public fora, such as conferences, workshops, and journals, can provide the necessary avenues for criticism of the research in conversational systems, just as they do in other sciences.

Author Contributions: Conceptualization, T.P.A.; Methodology, T.P.A.; Refining of Concept and Methology: F.L. and M.L.; Investigation, T.P.A.; Writing—Original Draft Preparation, T.P.A.; Writing—Review & Editing, F.L. and M.L.; Supervision, F.L. and M.L.

Funding: This research received no external funding.

Acknowledgments: The authors will like to appreciate and thank the Managing Editor and the anonymous reviewers of this essay for the very valuable feedback they provided.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
IR	Information Retrieval

References

1. Creath, R. Logical Empiricism. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2017.
2. Kuhn, T.S. *The Structure of Scientific Revolutions*; University of Chicago Press: Chicago, IL, USA, 2012.
3. Kazmier, L.J. *Theory and Problems of Business Statistics*; McGraw-Hill: New York, NY, USA, 2004.
4. Burgan, D. *Dialogue Systems and Dialogue Management*; Technical Report; DST Group Edinburgh: Edinburgh, Australia, 2016.
5. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
6. Thomason, R. Logic and Artificial Intelligence. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2018.
7. Indurkhya, N.; Damerau, F.J. *Handbook of Natural Language Processing*; CRC Press: Boca Raton, FL, USA, 2010; Volume 2.
8. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Boston, MA, USA, 2009.
9. Manning, C.D.; Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
10. Qiu, M.; Li, F.L.; Wang, S.; Gao, X.; Chen, Y.; Zhao, W.; Chen, H.; Huang, J.; Chu, W. Alime chat: A sequence to sequence and rerank based chatbot engine. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Short Papers; Volume 2, pp. 498–503.
11. Song, Y.; Yan, R.; Li, X.; Zhao, D.; Zhang, M. Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv* **2016**, arXiv:1610.07149.
12. Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.; Pineau, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
13. Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; Weston, J. Personalizing Dialogue Agents: I have a dog, do you have pets too? *arXiv* **2018**, arXiv:1801.07243.
14. Li, H. A Short Introduction to Learning to Rank. *IEICE Trans. Inf. Syst.* **2011**, *E94-D*, 1854–1862. [CrossRef]
15. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
16. Mutton, A.; Dras, M.; Wan, S.; Dale, R. GLEU: Automatic evaluation of sentence-level fluency. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 25–27 June 2007; pp. 344–351.
17. Clark, A.; Fox, C.; Lappin, S. *The Handbook of Computational Linguistics and Natural Language Processing*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
18. Belz, A.; Reiter, E. Comparing automatic and human evaluation of NLG systems. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 5–6 April 2006.
19. Thagard, P.R. Why astrology is a pseudoscience. *PSA Proc. Bienn. Meet. Philos. Sci. Assoc.* **1978**, *1978*, 223–234. [CrossRef]
20. Douven, I. Abduction. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2017.

21. Reiss, J.; Sprenger, J. Scientific Objectivity. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2017.
22. Longino, H. *Values and Objectivity*; Princeton University Press: Princeton, NJ, USA, 1998.
23. Keselj, V. *Speech and Language Processing*; Jurafsky, D., Martin, J.H., Eds.; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2009; ISBN 978-0-13-187321-6.
24. Popper, K. *Conjectures and Refutations: The Growth of Scientific Knowledge*; Routledge: Abingdon, UK, 2014.
25. Lipton, P. *Inference to the Best Explanation*; Routledge: Abingdon, UK, 2003.
26. Lakatos, I. Science and pseudoscience. *Philos. Pap.* **1978**, *1*, 1–7.