Unveiling Practices and Challenges of Machine Teachers of Customer Service Conversational Systems

Heloisa Candello
IBM Research
São Paulo, BR
hcandello@br.ibm.com

Mairieli Wessel
IBM Research
São Paulo, BR
mairieli@ibm.com

Claudio Pinhanez
IBM Research
São Paulo, BR
csantosp@br.ibm.com

Sara Vidon
IBM Research
São Paulo, BR
Sara.Vidon@ibm.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
This paper describes a set of qualitative interventions which aimed to unveil the practices of teaching conversational machines used in the automatic customer service. The study aimed to understand the activity of mapping information into conversational systems platforms to create chatbots (text or voice-based) to attend end-users in conjunction with call centers. We interviewed eleven domain experts with non-machine learning skills responsible for curating the content of the chatbots in two contexts. The first was in the domain of human resources and the second was in a banking domain. Additionally, we conducted four design workshops with experienced curators to understand deeper their challenges when teaching novice curators. We describe some of the fundamental tasks of content curators and we list a group of challenges and opportunities for improving the machine teacher's practices and supporting decision making.

Author Keywords
Content curation; conversational systems; knowledge representation; chatbots.

CCS Concepts
•Human-centered computing → **Human computer interaction (HCI);** User studies;

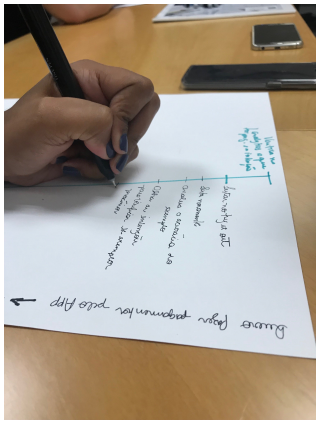

Figure 1: Timeline activity.

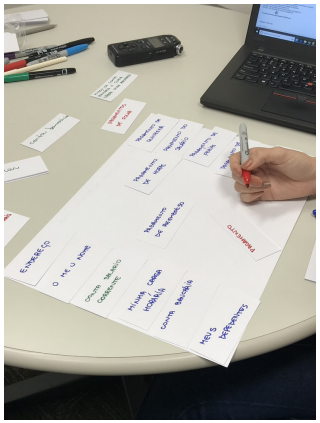

Figure 2: Explaining the timeline.

Introduction

Machine teaching is the artificial intelligence (AI) discipline which aims to make *machine learning (ML)* trainers more productive at building systems, through the use of high-level knowledge [4]. Additionally, machine teaching aims to give non-machine learning experts a more natural and simplified process of transferring information to the machine [1, 4]. We call *machine teachers* both types of professionals which are behind the everyday development and maintenance of ML systems.

The practices of machine teachers have not not been well studied, especially in professional contexts. These professionals aim to provide an enjoyable and effective experience for end-users of conversational systems in several customer care domains such as banking, healthcare, sales, IT help, and human resources. They often interact with technical specialists, who set the limitations of the conversational platforms and how the content should be included in the conversation flow. Most of the machine teachers play the role of content curators and are, in different degrees, domain experts, but often they are not specialists machine learning development itself.

In this study, we focus on machine teachers who act as a content curators and usually have IT specialists in machine learning to assist in their task. We conducted a set of semi-structured interviews and four workshops with content curators of high-end, professionally-built enterprise chatbots, to deep dive into their everyday practices. In this paper, we describe the methodology applied, their main practices, and the main challenges they face as a result of analyzing the data collected in those interventions. Our goal is to shed light on the practices and challenges of this emerging profession of machine teaching.

Initial Interviews with Machine Teachers

We applied the *Critical Decision Method* approach [2, 5] to unveil the practices and reasoning of the content curators from the interviews and workshops data. Participants were recruited by e-mail. Overall, we conducted 11 individual semi-structured interviews. Each interview took 45min-60min. Six participants worked in a human resources department of an IT company and the other five worked in a large scale bank. All of them were responsible to provide and maintain content to AI conversational systems which deliver customer care information to the end-user.

Methodology

First, we wanted to understand the mental process of the participants, asking them to describe a situation where re-training (teaching) of the system was needed as identified by the negative feedback of final users. They were requested to build a timeline of events while they explained procedures adopted, using *What IF questions* [5].

Results

The beginning of the timeline was the negative feedback provided by the user, as identified from call-logs, followed by the actions taken to identify and solve the problem, and finalizing with the updating of the training corpus.

This activity helped us to unveil the primary tools they adopt to give support to the conversational platform in use, the sources they access to create the answers and examples of end-user questions, and the validations required and their related collaborative tasks.

Design Workshop with Experienced Curators

We conducted a 4-day design workshop aimed at understanding challenges expert curators face when teaching novice ones and to explore tools which could facilitate their work. We focus here Four expert curators and four re-

searchers were invited to participate. Two expert curators work in chatbots for the auto industry, one for a bank, and another for a telecommunications company. During the workshop, researchers and expert curators were grouped in pairs to exchange their knowledge. Each workshop section lasts between 60 and 75 minutes.

Methodology

1. **Day 1: As-Is scenario map.** Groups were requested to draw an “as-is scenario map” representing everyday practices of the curators when teaching the machine. We provided the participants with the main phases of the curators’ timeline, from receiving negative feedback to including new content to the system. Each group chose a different context to work on, brainstorming individually what curators would be *doing*, *thinking* and *feeling* throughout the phases.

2. **Day 2: Stakeholder map and challenges identification.** First, an activity was conducted to identify the system stakeholders present in the four “as-is scenario maps” previously built by the groups. Then, one of the contexts was selected to be deeper explored. Based on that specific situation, all groups reported the main challenges which occur in each phase of the teaching process.

3. **Day 3: Taxonomy exploration.** In this workshop section, expert curators and researchers explored the taxonomy models of the information they use to organize the content to be taught to the machine and its associated challenges.

4. **Day 4: Interface proposal.** Finally, groups brainstormed the conception of a new interface to solve the challenges emerged in the previous sections. After, they were asked to draw the new interface prototype in six distinct steps.

Figure 3: Discussing the As-Is scenarios.

Results: Content Curatorial Practices

Curators might want to include or update information into the conversational platforms on several occasions. The usual case is to “map" new user intents which the machines might not be able to correctly identify. The curator knows those drawbacks by reviewing the logs of the conversations of users with the chatbot and by reviewing the end-users experience feedback forms. We focused on those practices and we identified a pattern on the sequence of tasks curators perform to solve mistakes and conflicts identified by the end-users.

Technological Context

Our participants use a platform that uses the paradigm of most of the conversational systems platforms built today, an *intent-action* approach [3]. The system is created by defining a basic set of user example questions, and the systems’ responses which should match to them. The term *intent* is adopted to describe the goal of a single group of example questions, so the essential task of the conversational platform is to identify the intent of a given question written or spoken by the user, and then output its associated answer or action.

In *user-initiative systems* (for example, typical QA systems), groups of questions from the user are mapped into a single answer from the systems, together with a set of variations. In *system-initiative* systems, the curators of the conversational systems have to provide sets of typical user questions for each output answer. Based on the intent of the question of the user, an action is produced often with the help of basic natural language parsing technology to help extract the system needed information. The AI system which determines the intent outputs a probabilistic "confidence score" interval before delivering an answer. The intent matching is often the most important source of problems in the develop-

ment of conversational systems due to the complexity and difficulty of analyzing natural language.

Some many different technologies and platforms can be used for intent matching. A common approach is to use a *template-based* system in which the intent is determined by the presence of manually defined terms or groups of words in the expected user questions. Template-based systems, although often the simplest way to start developing a conversational system, suffer from two key problems. First, it is hard to capture the many nuances of human language. Second, it is challenging to track the source of errors and debug the system successfully. The content curators we interviewed explained to us those two challenges in detail as described below.

Figure 4: Identification of challenges during the day 2 of the workshop.

Curator Steps to Improve Content
We identified a sequence of practices curators adopt to teach machines to better respond to end-user utterances.

1. **Review the call logs:** Some of curators receive 10,000 utterances per day to review. They usually choose a sample to analyse.

2. **Differential diagnostic:** After choosing an example or a set of examples, they try determine whether the problem is caused by a conflict of intents or whether there is a need of a new intent.

3. **If conflict:** Curators test the user question selected from the logs and identify the confidence score of the intent that the system chose to match to the utterance of the user. They then review the examples of the selected intent and compare to the examples of the intent that the system should have selected instead. To correct it, they edit, add, and delete examples which might be confounding the AI system (e.g. the structure of the utterance, similar questions, or lack of relevant examples) and, most importantly, map the exact user question selected in the logs to the suitable intent as an example question. Following, curators try out the example question and analyse the output level of confidence. Moreover, they also test the result of the ML training in the end-user interface to make sure it is correct.

4. **If new intent:** Some topics of end-user questions might be new to the corpus. In this case, curators select end-user questions of the new topics and use them as example questions to create new intents in the database. They have to give a name, an ID to the intent, and often use a taxonomy to construct the ID. Using the taxonomy will help them and peers to find the intent in the future. From the group of end-user questions selected they choose one question, prototypical of the situation, called the *canonical question*, that will represent that group of example questions of the new intent. After that, they identify the sources which are needed to provide the information to extract and create an answer. The answer is reviewed by the *product owners*, specialists in the content and proofread. Often, the new intent and its answer must have management approval to be included in the corpus. The answer must have a standard language or personality according to the identity guidelines of the conversational system. After all the approvals are secured, some curators add the answer to the system, while others use spreadsheets to send the answers and example questions created to a developer.

Some curators have access to the full platform features and can update and solve conflicts or create new intent themselves. Others have supporting tools that help them to see the level of confidence if the utterances although they

cannot edit direct into the platform, as a consequence they send their updates to a developer to update the system. We also identify curators that do not have any access to supporting tools and platforms. Those contact the developers by e-mail and explain the errors identified in the logs and in the end-user platforms. The developers analyse the situation and ask curators to create new examples or answers in a spreadsheet for them to update the system.

In the case of new intents, after it is incorporated to the corpus by the curators or the developers, the curator tryout the example questions and check the outcome level of confidence. If there is any divergence, they start the process again to solve any conflicts that might arise. If not, the developer add those to the system, and advise by e-mail the curators for them to test in the end-user platform. In case of any divergences, the curator contacts the developer and send the spreadsheet again for update the system.

Figure 5: Drawing the new interface with 6 steps.

Main Challenges Faced by Curators
Expert curators do several manual tasks which impact the productivity and the quality of the end-result, and therefore of the user experience. We describe the most relevant ones.

1. Identifying new topics in millions of questions.
2. Managing concurrent incompatible software system and permissions.
3. Using several systems to create content in the same time.
4. Determining the priorities of the issues to fix.
5. Mapping to each other similar dialogue failures.
6. Handling several stakeholders that had to approve new content.
7. Fixing duplicate information in the corpus created by different curators.
8. Using several sources of content to extract answers.
9. Reusing similar content.
10. Avoiding intent IDs which might confound the curators.
11. Creating mechanisms for automatic updates for similar categories.
12. Dealing with flow digression, disambiguation, and lack of content.
13. Handling diverse taxonomy intent IDs for multiple clients.
14. Reusing the corpus in another language.
15. Tracking editions by each curator.

In practice, in the context of limited supporting tools, curators must trust their own experience and familiarity with the content to guarantee a good user experience with the chatbot.

Final Remarks

We see this paper as an initial, preliminary work to understanding the practices and needs of the content curators of conversational systems. We believe it is a starting point to uncover the fundamental work curators do today to create and maintain complex conversational systems, including many of the chatbots being used by enterprises in different parts of the world.

We acknowledge the need to explore new methodologies to be applied in this problem, and of deeper and larger scale studies. We are particularly curious to know which other methods, tools, and practices other curators of AI content have, especially in other domains.

REFERENCES

[1] Tasneem Kaochar, Raquel Torres Peralta, Clayton T Morrison, Ian R Fasel, Thomas J Walsh, and Paul R Cohen. 2011. Towards understanding how humans teach robots. In *International Conference on User*

Modeling, Adaptation, and Personalization. Springer, 347–352.

[2] Gary A Klein, Roberta Calderwood, and Donald Macgregor. 1989. Critical decision method for eliciting knowledge. *IEEE Transactions on systems, man, and cybernetics* 19, 3 (1989), 462–472.

[3] Jetze Schuurmans and Flavius Frasincar. 2019. Intent Classification for Dialogue Utterances. *IEEE Intelligent Systems* (2019).

[4] Patrice Y. Simard, Saleema Amershi, David Maxwell Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Robert Wernsing. 2017. Machine Teaching: A New Paradigm for Building Machine Learning Systems. *CoRR* abs/1707.06742 (2017).

[5] Hazel Taylor. 2005. A critical decision interview approach to capturing tacit knowledge: Principles and application. *International Journal of Knowledge Management (IJKM)* 1, 3 (2005), 25–39.