10-9-2023

Towards Designing a NLU Model Improvement System for Customer Service Chatbots

Daniel Schloß
Karlsruhe Institute of Technology, Germany, daniel.schloss@kit.edu

Ulrich Gnewuch
Karlsruhe Institute of Technology, Germany, ulrich.gnewuch@kit.edu

Alexander Maedche
Karlsruhe Institute of Technology, Germany, alexander.maedche@kit.edu

Towards Designing a NLU Model Improvement System for Customer Service Chatbots

Research in Progress

Daniel Schloß[1], Ulrich Gnewuch[1] and Alexander Maedche[1]

[1] Karlsruhe Institute of Technology, human-centered systems lab, Karlsruhe, Germany
{daniel.schloss,ulrich.gnewuch,alexander.maedche}@kit.edu

Abstract. Current customer service chatbots often struggle to meet customer expectations. One reason is that despite advances in artificial intelligence (AI), the natural language understanding (NLU) capabilities of chatbots are often far from perfect. In order to improve them, chatbot managers need to make informed decisions and continuously adapt the chatbot's NLU model to the specific topics and expressions used by customers. Customer-chatbot interaction data is an excellent source of information for these adjustments because customer messages contain specific topics and linguistic expressions representing the domain of the customer service chatbot. However, extracting insights from such data to improve the chatbot's NLU, its architecture, and ultimately the conversational experience requires appropriate systems and methods, which are currently lacking. Therefore, we conduct a design science research project to develop a novel artifact based on chatbot interaction data that supports NLU improvement.

Keywords: Customer Service, Chatbots, Analytics, Language Understanding

1 Introduction

Chatbot technology is now widely used in customer service to answer customer questions or handle concerns end-to-end (Schuetzler et al., 2021).

In the rather short conversations with customer service chatbots, customers place particular value on performance (Diederich et al., 2022; Følstad and Skjuve, 2019). However, as known from research and practice, many customers still have negative experiences with chatbots. This is often due to weaknesses in the chatbots' natural language understanding (NLU). In the worst case, customers experience a breakdown ("Sorry, I did not understand that") or a mismatch (an incorrect or inappropriate response) (Benner et al., 2021; Følstad and Taylor, 2021). To avoid this, various NLU improvements performed by chatbot managers are possible: The language model of a customer service chatbot can be supplemented with additional topics or training data. Additionally, the architecture of the NLU can be adapted or linguistic fine tuning can be performed (AI Multiple, 2022). However, with growing options in NLU training, chatbot managers must make very thoughtful decisions as these can affect the customer service experience of possibly thousands of customers. However, these decisions do not have to be made uninformed. As research as well as practical advice on chatbot and NLU development suggest, real-world interaction data is valuable for a "conversation-driven development" (Akhtar et al., 2019; Rasa, 2023). By "reviewing conversations on a regular basis" (Rasa, 2023), improvements, especially with regard to NLU and Dialog Management (DM), can be discovered and conducted (Følstad and Taylor, 2021). However, it is still largely unknown how a system providing such insights based on real-world conversation data needs to be designed to enable the discovery of such improvement opportunities on a mass large scale and in a more automated manner (Chen and Beaver, 2022). To address this gap, we formulate the following research question:

RQ: *How ought a system for NLU model improvement of customer service chatbots to be designed?*

Accordingly, the goal of the project presented in this paper is to generate (instantiated) design knowledge for a user-centric and usage data-based system for NLU model improvement. Regarding the usage data, we aim to provide a high degree of automation (compared to manual analytical approaches, e.g. Kvale et al., 2020). Regarding the users of the system, we reveal how and what descriptive as well as prescriptive insights into conversation data help NLU designers and chatbot managers on an aggregate as well as detailed level to improve their NLU (and DM). Additionally, to ensure practical relevance and user-centeredness, this research project is conducted as design science research (DSR) project in collaboration with an industry partner. The setup is presented in chapter three. Prior to this, we introduce and build on related research from the fields of chatbot analytics and communication theory. In chapter four we provide a first insight into our project and system. Chapter five concludes with a summary and outlook.

2 Related Work: Chatbots in Customer Service

In customer service, task-focused chatbots are used (Schuetzler et al., 2021). With these, customers typically have short conversations regarding specific concerns where they expect to be served fast and effectively (Diederich et al., 2022; Følstad and Skjuve, 2019). They answer questions, e.g. via texts or links or start guided dialogs in which multiple conversational "turns" takes place, for example, when asking prescripted questions to identify a user to ultimately process a business transaction.

While open domain chatbots like ChatGPT or Replika build on non-deterministic Natural Language Generation (NLG), the architecture of customer service chatbots is intent-based at its core (OpenAI, 2023; Replika, 2023; Schuetzler et al., 2021). As **Figure 1** shows, customer utterances are classified to an intent with a probability by a Natural Language Understanding (NLU) system. Next, the dialog management (DM) controls the chatbots' actions based on the NLU information and defined logical rules. For customer service chatbots, the responses are then typically retrieved from a knowledge base (KB) (Kucherbaev et al., 2018). This deterministic approach has the advantage that the responses, which only need to cover certain topics, are fully controllable for service managers as opposed to probalistic NLG responses (Galitsky, 2019).

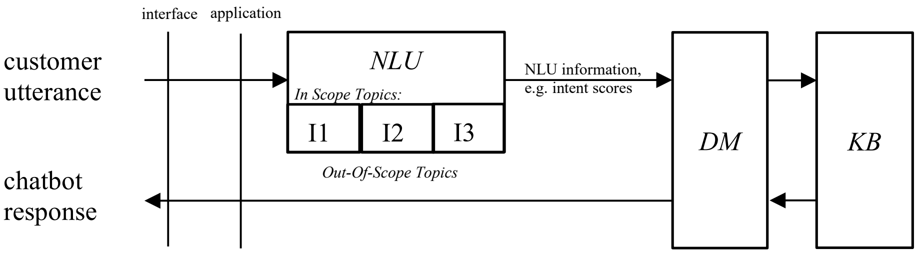


Figure 1. Typical Customer Service Chatbot Architecture (Adamopoulou et al., 2020)

However, problems can arise with this language-based classification problem. In chatbot research, these are summarized as conversational or chatbot "breakdowns," i.e. situations in which the chatbot responds with "Sorry, I can't help you." (Benner et al. 2021). A large body of research has been devoted to the moderation and resolution of these situations within a conversation (e.g., Ashktorab et al., 2019), as they cause customer dissatisfaction and poor reputation for the technology and the company (Schuetzler et al., 2021; van der Goot et al., 2021). A breakdown is caused by the failure to succesfully detect a correct intent. In addition, an incorrect response, a so-called (intent) mismatch can occur as well (or false positive, Følstad and Taylor, 2021). Both types of errors must be corrected by a chatbot manager. This often requires manual review of conversations, as well as small adjustments within a large and difficult-to-manage set of intents, training data, and features (Beaver and Mueen, 2021). Large NLU systems, e.g., IBM Watson or Microsoft LUIS, do provide rudimentary reporting for NLU revision (Diederich et al., 2019). However, NLU improvement support is often based only on internal training data omitting detailed analysis of the real-world conversation data and specific interpretation capabilities. For this reason, user-friendly chatbot analytics systems can help to better understand conversational, specifically NLU, problems and suggest appropriate actions (Beaver and Mueen, 2020; Yaeli and Zeltyn, 2021).

3 Research Method

To address the problem of detecting and adequately resolving NLU problems in customer service chatbots, we are designing a data-driven NLU model improvement system following the design science research (DSR) paradigm (Gregor and Jones, 2007). To ensure the relevance of the project besides scientific rigor we collaborate with an industry partner (Hevner, 2007). Our industry partner is an IT and service provider for the utility industry, operating and maintaining customer service chatbots for over 25 small to very large utilities. The DSR project, shown in Table 1, follows the established structure of Kuechler and Vaishnavi (2008) and consists of two design cycles. The first design cycle (DC1) is dedicated to the descriptive analysis of chatbot log data (user messages and internal metrics). To this end, we first explored the problem space building on research on customer service chatbots and NLU. In the suggestion phase, we explored insights from (chatbot) analytics and communication research to outline methods for log data analysis for NLU optimization. Moreover, we had access to more

600.000 user messages and NLU log data. Based on this, we have drafted a first system and are currently in the development phase. Currently, it mainly contains descriptive exploration of chat data for NLU improvement, e.g. filter functions (e.g. according to intent scores) and corresponding visualizations. In the second step (DC2), we plan to upgrade the system for prescriptive analysis. By connecting the support system not only to the log data source but to the NLU system, we finally want to provide usage data-driven decision support and recommendations for the NLU, e.g. regarding training data. This will be guided by two evaluations we plan in both design cycles, first a think aloud study with experts of the industry partner, in DC2 an evaluation including external experts such as researchers and a case study where we evaluate the NLU performance, e.g. classification, ex ante and ex post implementing changes suggested by our system.

Table 1. Overview of the Design Science Research Project

	DC1: Log Data Analysis	DC2: Log – Training Data comparison
1) Problem Awareness	Chatbot and NLU Research	
2) Suggestion	Theory and Sample Dataset	Concept for Data Integration of System
3) Developement	System Prototype	Update: NLU/KB Data Integration
4) Evaluation	Expert Think-Aloud Sessions	Expert Assessment of System Guidance
5) Conclusion	Refinement and Planning DC2	Summary of Design Knowledge

4 NLU Model Improvement System

4.1 Problem Awareness

If the NLU of a chatbot does not understand the customer's concern and the chatbot does not provide a relevant answer, it creates customer frustration and runs counter to the goal of automating customer service (Schuetzler et al., 2021; van der Goot et al., 2021). Therefore, chatbot managers need to continuously improve the NLU. But just like Large Language Models, intent-based NLU systems are often referred to as a “black box”. Chatbot managers have only limited insight into the statistical operations of the NLU offered by the chatbot platforms (Diederich et al., 2019; Schuurmans and Frasincar, 2019). What they can control, is the selection and partitioning of training data, as well as the use of specific NLU features (Ruane et al., 2020). Since language understanding in intent-based (retrieval) chatbots is based on a classification problem, the performance of the model is often evaluated by means of precision, recall or F1 score, testing training utterances internally in the NLU (Rasa, 2023). Against this background, there are different reasons why breakdowns (or mismatches) occur. First, it may be that a customer has simply entered (1) semantic nonsense (e.g., “Colorless green ideas sleep furiously,” Chomsky and Lightfoot (2007)) or (2) grammatical nonsense (e.g., “sadsasdsde”) the NLU does not need cover. However, it may also be that he or she expresses appropriately, but the topic addressed is (3) out-of-scope of the chatbot (Følstad and Taylor, 2021). In all cases, the actual tokens (e.g. “sleep”) or the semantic meaning will intentionally be underrepresented in the NLU training data (Debortoli et al., 2016). It is also possible, as we observed in the provided log data, that customers use familiar words, but (4) formulate too short or verbose (Beaver and Mueen, 2020; Reinkemeier and Gnewuch, 2022). If all that is not the case and a breakdown occurred, it might be that (5a) the training data of the NLU may not be grammatically (i.e. syntactically) sufficient (e.g., “I want to cancel” vs. “Cancellation please) or (5b) lacks synonyms for equal semantics (e.g., “I want to buy a product” vs. “I want to purchase a product”) (Jurafsky and Martin, 2009; Sperber and Wilson, 1986). In this case, chatbot managers need to add on the training data. Last, the NLU itself may have (6) supervised learning performance issues such as imbalances (over-/underfitting) or poorly weighted linguistic features (Rasa, 2023). All of these causes ultimately lead to poor user experiences, but can be mitigated incrementally through informed NLU improvements, particularly based on real-world usage data (Diederich et al., 2019).

4.2 Suggestion

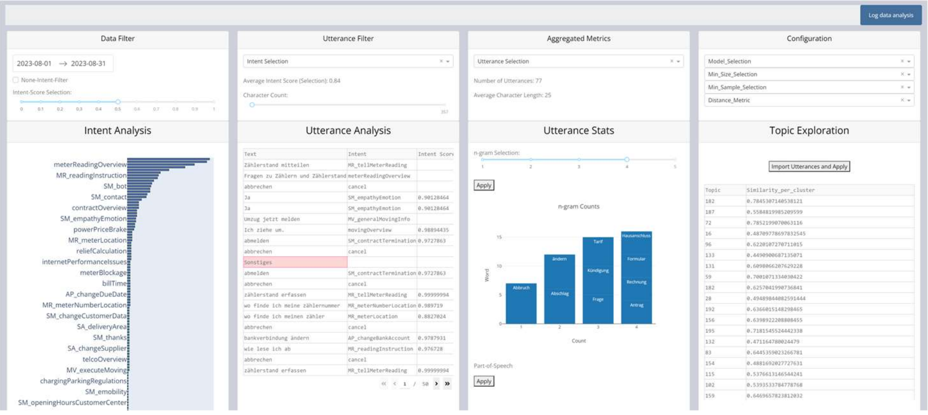
What can chatbot managers do to improve their language model to meet customer needs? To inform the design of our improvement system, we draw on the communication theories “Grounding in Communication” and “Communication Accommodation Theory” (Clark, 1996; Giles and Ogay, 2007). These theories propose the main components of communication:

Semantical: First, successful communication is characterized by all parties having a “common ground”, i.e. a shared level of knowledge of the world (Clark, 1996). With regard to customer service chatbots, customers expect a broad knowledge base on all their concerns. Therefore an important component of the NLU Model Improvement System is the discovery of the situations where the chatbot violates the assumed common ground, see (3). These situations can be used as opportunities to create new intents and training data (Chen and Beaver, 2022). We have addressed this in the prototype shown on the next page by means of filter functions for unrecognized utterances and are also working on fast topic modeling for an easier overview of the unknown topics.

Syntactical: Second, according to “communication accommodation” good communication is characterized by the fact that a convergence of communication parties happens in a conversation. This can refer to social aspects or gestures, concerning text-based chatbots as in this stage of the project, this convergence needs to manifest in the applied language (understanding), see (4), (5a), (5b) (Giles and Ogay, 2007). Customer-friendly customer service chatbots must therefore not only choose their answers in the style of the customer, they also need to acquire the customer's understanding of the language (semantic and grammatical) over time. For our proposed system, this means that it must provide insight into the patterns of customers, especially with regard to utterances that are not yet well understood. This is represented by our initial filters, but also by linguistic metrics such as most frequent words or part-of-speech.

Structural: Third, there are internal technical premises of NLU systems that affect their performance independently of the specific semantics or syntax of the chosen training data, see (6). An example of this is a “skewed” distribution of training data leading to imbalances (Ruane et al., 2020). We will address the analysis of these structural NLU improvement opportunities in DC2 when the NLU system is connected to our system.

4.3 Development



2023-08-01 → 2023-08-31

□ None-Intent-Filter

Intent-Score Selection:

Intent Selection

Average Intent Score (Selection): 0.84

Character Count:

Utterance Selection

Number of Utterances: 77

Average Character Length: 25

Model_Selection

Min_Size_Selection

Min_Sample_Selection

Distance_Metric

Text	Intent	Intent Score
Zählerstand mitteilen	MR_tellMeterReading	
Fragen zu Zählern und Zählerstand	meterReadingOverview	
abbrechen	cancel	
Ja	SM_empathyEmotion	0.90128464
Ja	SM_empathyEmotion	0.90128464
Umzug jetzt melden	MV_generalMovingInfo	
Ich ziehe um.	movingOverview	0.98894435
abmelden	SM_contractTermination	0.9727863
abbrechen	cancel	
Sonstiges		
abmelden	SM_contractTermination	0.9727863
abbrechen	cancel	
zählerstand erfassen	MR_tellMeterReading	0.99999994
wo finde ich meine zählernummer	MR_meterNumberLocation	0.989719
wo finde ich meinen zähler	MR_meterLocation	0.8827024
abbrechen	cancel	
bankverbindung ändern	AP_changeBankAccount	0.9787931
wie lese ich ab	MR_readingInstruction	0.976728
abbrechen	cancel	
zählerstand erfassen	MR_tellMeterReading	0.99999994

n-gram Selection:

Part-of-Speech

Topic	Similarity_per_cluster
182	0.7845307140538121
187	0.5584819985209599
72	0.7852199070063116
16	0.48709778697832545
96	0.6220107270711015
133	0.4490900687135071
131	0.6098066207629228
59	0.7001071334030422
182	0.6257041990736841
28	0.49489844082591444
192	0.6366015148298465
156	0.6398922208808455
195	0.7181545524442338
132	0.471164780024479
83	0.6445359023266781
154	0.4881692027727631
115	0.5376613146544241
102	0.5393533784778768
159	0.6469657823812032

Figure 2. Screenshot of the NLU Model Improvement System Prototype

The prototype of the NLU improvement system, developed on the basis of python modules and connected to a MySQL log database, is built in a top-down design. This allows for generous filtering and mining of many customer utterances to apply detailed analysis for specific utterances (Beaver and Mueen, 2020). First, timestamps can be selected (due to NLU adaptations), but also intents can be filtered (left). In addition, the tool provides an overview on current intent utilization, since a too "long tail of intents" can also be cause of structural NLU performance problems (Greyling, 2022). Moreover, chatbot managers have the possibility of selecting the most interesting utterances via intent scores/thresholds or the text length as the first linguistic feature. After that, they can examine in-depth structural linguistic metrics of their selection or individual utterances (e.g. n-grams or Part-of-Speech, Manning et al., 2014). In DC2, it is planned to expand these metrics and compare them to NLU training data according to the customer service chatbot experts' input and requirements. Finally, the right-hand side of the system offers the possibility to apply parameterizable topic mining to an utterance selection. In this way, new (in-scope) topics can be discovered (Chen and Beaver, 2022).

5 Conclusion and Outlook

The NLU model improvement system we present addresses the current problem of imperformant customer service chatbots. We are planning the technical completion of the system in order to evaluate it with customer service experts in the next DSR stage. The descriptive and exploratory nature of the system, as applied to log data, will then be augmented with research and practical expertise to provide a stronger guiding function.

References

Akhtar, M., Neidhardt, J., and Werthner, H. (2019). The potential of chatbots: analysis of chatbot conversations. In 2019 IEEE 21st conference on business informatics (CBI) (Vol. 1, pp. 397-404). IEEE.

AI Multiple, (2022). In-Depth Guide Into Chatbots Intents Recognition in 2023. https://research.aimultiple.com/chatbot-intent/, Accessed: 08.03.2023.

Beaver, I., and Mueen, A. (2020). Automated conversation review to surface virtual assistant misunderstandings: Reducing cost and increasing privacy. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 08, pp. 13140-13147).

Chen, X., and Beaver, I. (2022). An Adaptive Deep Clustering Pipeline to Inform Text Labeling at Scale. Baltimore, Maryland, USA: Proceedings of the 39 th International Conference on Machine Learning.

Chomsky, N., and Lightfoot, D. W. (2002). Syntactic structures. Walter de Gruyter.

Clark, H. H. (1996). Using language, Cambridge: Cambridge University Press.

Diederich, S., Brendel, A. B., and Kolbe, L. M. (2019). Towards a Taxonomy of Platforms for Conversational Agent Design. Internationale Tagung Wirtschaftsinformatik, Siegen, Germany.

Diederich, S., Brendel, A. B., Morana, S., and Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. Journal of the Association for Information Systems, 23(1), 96-138.

Debortoli, S., Müller, O., Junglas, I., and Vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. Communications of the Association for Information Systems, 39(1), 7.

Følstad, A., & Skjuve, M. (2019, August). Chatbots for customer service: user experience and motivation. In Proceedings of the 1st international conference on conversational user interfaces (pp. 1-9).

Følstad, A., and Taylor, C. (2021). Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. Quality and User Experience, 6(1), 6.

Galitsky, B. (2019). Developing enterprise chatbots. New York: Springer International Publishing.

Giles, H., and Ogay, T. (2007). Communication accommodation theory.

Gregor, S., and Jones, D. (2007). The Anatomy of a Design Theory. Journal of the Association for Information Systems, 8(5), 312–335.

Greyling, C. (2022). Solving for the long tail of intent distribution. https://cobusgreyling.medium.com/solving-for-the-long-tail-of-intent-distribution-24daa372fcc, Accessed: 08.03.2023

Greyling, C. (2023). NLU & NLG Should Go Hand-in-Hand. https://cobusgreyling.medium.com/nlu-nlg-should-go-hand-in-hand-8cc7952ebab8, Accessed: 08.03.2023.

Hevner, A. R. (2007). A three cycle view of design science research. Scandinavian journal of information systems, 19(2), 4.

Jurafsky, D. and Martin, J. H. (2009). Speech and Language Processing, 2nd Edition, New Jersey: Prentice-Hall.

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: anatomy of a research project. European Journal of Information Systems, 17(5), 489-504.

Kvale, K., Sell, O. A., Hodnebrog, S., & Følstad, A. (2020). Improving conversations: lessons learnt from manual analysis of chatbot dialogues. In Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3 (pp. 187-200). Springer International Publishing.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60..

Open AI (2023). ChatGPT. https://chat.openai.com/chat, Accessed: 08.03.2023.

Rasa (2023). Conversation-Driven Development. https://rasa.com/docs/ , Accessed: 08.03.2023.

Reinkemeier, F., and Gnewuch, U. (2022). Designing effective conversational repair strategies for chatbots. ECIS 2022 Research Papers

Replika (2023), The AI companion who cares. https://replika.ai/, Accessed: 08.03.2023

Ruane, E., Young, R., & Ventresque, A. (2020, March). Training a chatbot with Microsoft LUIS: effect of intent imbalance on prediction accuracy. In Proceedings of the 25th International Conference on Intelligent User Interfaces Companion (pp. 63-64).

Schuetzler, R. M., Grimes, G. M., Giboney, J. S., and Rosser, H. K. (2021). “Deciding Whether and How to Deploy Chatbots”, MIS Quarterly Executive 20 (1), 1-15.

Schuurmans, J., and Frasincar, F. (2019). Intent classification for dialogue utterances. IEEE Intelligent Systems, 35(1), 82-88.

Sperber, D. and Wilson, D. (1986). Relevance: Communication and cognition, Vol. 142. Cambridge, MA: Harvard University Press.

Van der Goot, M. J., Hafkamp, L., and Dankfort, Z. (2021). Customer service chatbots: A qualitative interview study into the communication journey of customers. In Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4 (pp. 190-204). Springer International Publishing.