
Proceedings Paper:
Aftab, Haris orcid.org/0000-0001-7981-1743, Hammad Hussain Shah, Syed and Habli, Ibrahim orcid.org/0000-0003-2736-8238 (2021) Classification of Failures in the Perception of Conversational Agents (CAs) and their Implications on Patient Safety. In: Mantas, John, Stoicu-Tivadar, Lăcrămioara, Chronaki, Catherine Chronaki, Hasman, Arie, Weber, Patrick, Gallos, Parisis, Crişan-Vida, Mihaela, Zoulias, Emmanouil and Sorina Chirila, Oana, (eds.) Public Health and Informatics:Proceedings of MIE 2021. Studies in Health Technology and Informatics. IOS Press, pp. 659-663.

https://doi.org/10.3233/shti210253

Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210253

Classification of Failures in the Perception of Conversational Agents (CAs) and Their Implications on Patient Safety

Haris AFTAB[a,1], Syed Hammad Hussain SHAH[b] and Ibrahim HABLI[a]
[a] *Department of Computer Science, University of York, York, United Kingdom*
[b] *Department of ICT and Natural Sciences, Norwegian University of Science and Technology, Aalesund, Norway*

Abstract. The use of Conversational agents (CAs) in healthcare is an emerging field. These CAs seem to be effective in accomplishing administrative tasks, e.g. providing locations of care facilities and scheduling appointments. Modern CAs use machine learning (ML) to recognize, understand and generate a response. Given the criticality of many healthcare settings, ML and other component errors may result in CA failures and may cause adverse effects on patients. Therefore, in-depth assurance is required before the deployment of ML in critical clinical applications, e.g. management of medication dose or medical diagnosis. CA safety issues could arise due to diverse causes, e.g. related to user interactions, environmental factors and ML errors. In this paper, we classify failures of perception (recognition and understanding) of CAs and their sources. We also present a case study of a CA used for calculating insulin dose for gestational diabetes mellitus (GDM) patients. We then correlate identified perception failures of CAs to potential scenarios that might compromise patient safety.

Keywords. Conversational Agents, Healthcare, Machine Learning, Patient Safety

1. Introduction

Conversational agents (CAs) are software programs that interact with users in natural language [1]. Some of the well-known commercially available CAs are Google Assistant, Apple Siri, and Amazon Alexa. There are two main types: task-oriented (or goal-oriented) [2] and open-domain CAs (or chatbots) [3]. The former aims to assist users in achieving a task or a goal in a specific domain while the latter focuses on maximising the user's engagement. Task-oriented CAs typically use a structured ontology that represents the knowledge source of their intended tasks.

The architecture of CAs typically consists of various components connected in a pipeline as shown in Figure 1. The perception of a CA refers to recognition handled by Automatic Speech Recognition (ASR) and Spoken/Natural Language Understanding (SLU/NLU) components. In this paper, we use the term '*failure*' to describe ways in which a CA might fail to complete a user's task and '*error*' refers to contributory factors which ultimately cause these failures. We identify failures in clinical CAs with special

[1] Corresponding author, Haris Aftab, Department of Computer Science, University of York, York, YO10 5GH, United Kingdom; E-mail: haris.aftab@york.ac.uk.

focus on the perception (recognition and understanding) in task-oriented CAs. Henceforth, the term 'CAs' thus refers specifically to task-oriented CAs.

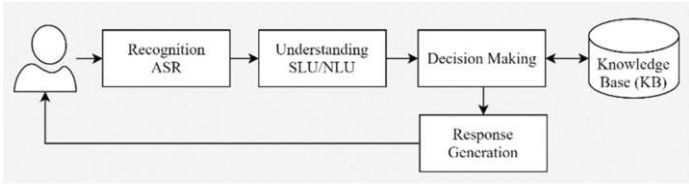

Figure 1. Architecture of a CA

CAs have been used in various industries such as e-commerce, travel, and sports [4]. A growing interest has been seen in their use in healthcare because of their potential use in providing 24-hour medical guidance, connecting patients to healthcare providers, helping clinicians in decision making etc. Common uses of CAs for patients include symptom checking, chronic disease management [5], health monitoring and medication adherence [6].

Although the technology behind CAs such as ASR and SLU/NLU has significantly improved, CA failures can affect the health condition of many patients at once. Previous studies show that the majority of failures in CAs come from weaknesses in their recognition [7, 8] and understanding [9] which ultimately influence their decision making process. In this work, we aim to explore and classify some of the failures that arise from the perception of CAs and their causes. We present a case study on a CA which calculates the insulin dose for patients with gestational diabetes mellitus (GDM). GDM occurs in women during pregnancy which increases their blood sugar levels and may threaten the life of both mother and the baby [10]. There are multiple ways to maintain blood glucose levels for the treatment of GDM, i.e. healthy diet, exercise, and medication (tablets or insulin injections). Here, we focused on the calculation of insulin dose through injections through a CA. We then investigate potential scenarios where specific CA failures might promise patient safety.

2. Method

We performed a literature search to find CAs and their applications in healthcare along with the taxonomy and classification of failures using search query given below: *((”conversational OR virtual OR digital OR smart”) AND (”agent OR assistant”) OR (”chatbot OR chatterbox”)) AND (”error OR failure OR issues OR faults OR safety”) AND (”medical OR ”healthcare” OR hospital OR medicine OR homecare”).*

We functionally classified failures in the perception of CAs based on relevant components in the architecture which is shown in Figure 1. We selected relevant studies which focus on the recognition (ASR) and understanding (NLU) failures in CAs.

CAs are widely used for managing chronic conditions such as diabetes [11]. Therefore, for our case study, we chose a critical CA that is used for calculating insulin dose in the treatment of GDM. We then analyse possible scenarios in insulin dose calculation by our CA. A wrong dose suggestion can be life-threatening for some patients.

3. Results

Recognition in CAs is achieved through ASR which transcribes speech input from the user and then apply knowledge from its vocabulary to correctly recognize the input text. Errors in speech transcribing can be caused by background noise induced into speech signal and misrepresentation of speech samples in the ML acoustic model of ASR. Non-native speakers, people with accents, elders, or children are some of the examples of input samples used for training an acoustic model [12]. Failure to include enough training examples for this ML model can result in inaccurate speech transcribing. CAs might fail to recognize an input due to the lack of vocabulary in the ML model of ASR component. The ASR tries to substitute words from the vocabulary it possesses and thus cause errors [7]. This failure may also occur when a user provides out of vocabulary input which the system was not designed to handle.

Table 1. Classification of failures in the perception of CAs

Component	Failure Class	Cause
Recognition	Inaccurate speech transcribing	Misrepresentation of speech samples in the acoustic model, and background noise
	Misrecognition of words	Incomplete vocabulary in ASR lexicon, and out-of-vocabulary user input
Understanding	Misunderstanding utterance	Incorrect slot-tagging, incomplete training examples, and ASR errors
	Non-understanding utterance	Out of vocabulary words, out of application utterance, and ASR errors

Understanding in a CA is performed by the SLU/NLU. The main failure classes are misunderstanding [13] and non-understanding utterance [14]. Misunderstanding utterance occurs when the system fails to get the correct semantic interpretation of the user input. Failure to understand the user intent mainly comes from insufficient training examples in ML model for intent classification [15]. Another cause of misunderstanding occurs due to incorrect slot-tagging [16] and ASR errors [17]. The non-understanding utterance failure occurs when a user asks a query which the system does not support. Using out of vocabulary words, out of application utterance, and ASR errors are the main causes of this error [14]. The results from the applied method for classification of perception errors are shown in Table 1.

Table 2 provides example scenarios from our case study and their correlation to identified failure classes (Table 1). The dose calculation method discussed in [18] was considered while developing the case study. Failures in recognition may cause inaccurate insulin dosage calculation (*incorrect transcription 116 instead of 160*) or type of insulin (*basal instead of bolus*). The examples given above show that the system might be susceptible to an unsafe response because of these failures. Similarly, failures due to misunderstanding may also occur where a CA might misunderstand the user input and provide an altogether different response (see Table 2). Asking how to inject insulin from CA might result in a non-understanding error in our example case study. The presented case study highlights the clinical CA failures and the need for the safety of such systems.

Table 2. Anticipated scenarios in the perception of CA failures for insulin dose management

CA Failure Class	Example Scenarios
Inaccurate speech transcribing	U: I want to calculate my insulin dosage and I am in my third trimester B: What is your weight in pounds? U: ***160*** B: Your insulin requirement is (***116***/2.2*0.9) 47.45 units
Misrecognition of words	U: What is my daily ***basal*** insulin dosage? B: Your daily ***bolus*** insulin dosage provided by rapid-acting insulin is…
Misunderstanding utterance	U: I want to ***calculate my insulin dosage*** and I am in my first trimester B: What is your weight in pounds? U: 120 B: Your ***weight in kilograms*** is (120/2.2) 54.43
Non-understanding utterance	U: What is the correct way to inject insulin in my body? B: Sorry, ***I do not understand***

4. Discussion

The current work presents a classification of failures in CAs based on the perception functions in its pipeline. Our approach is beneficial for analyzing potential safety hazards and risks pertaining to these agents. Other classification schemes examine failures in CAs based on performance measures [14, 19, 20] and do not give insight into the safety implications. The work presented by O'Halloran et. al. [21] inspired our method, which also focuses on finding hazards in a functional and systematic manner. In addition to that work, we mapped our method to real-world scenarios for the demonstration of possible safety failures in CAs. It can be seen from the scenarios shown in Table 2 that even in the simplest application a user might receive unsafe clinical advice.

5. Conclusions

In this paper, we provided an approach to classify failures in CAs and presented a case study to correlate these failure classes to real-world scenarios. Failures in CAs are most likely to arise from the user interaction, and errors in ML models and other components. Our case study shows how these perception failures can influence its decision making. As such, before deployment of CAs in a safety-critical environment, it is imperative to consider such failures.

The limitation of our proposed work is that it is based on a preliminary review in a fast-moving field and has only considered the perception and understanding components. We next plan to implement the CA discussed in present case study and investigate in detailed and more thoroughly the safety-related failures based on our method. We then aim to extract detailed safety requirements for that CA and determine design measures to mitigate the patient safety risks associated with these failures.

Acknowledgement

This research has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020 under Grant Agreement No. 812.788.

References

[1] Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, Surian D, Gallego B, Magrabi F, Lau AYS, Coiera E. Conversational agents in healthcare: A systematic review. J Am Med Informatics Assoc. 2018;25(9):1248–58.
[2] Gao J, Galley M, Li L. Neural approaches to conversational ai. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018. p. 1371–4.
[3] Harms J-G, Kucherbaev P, Bozzon A, Houben G-J. Approaches for dialog management in conversational agents. IEEE Internet Comput. 2018;23(2):13–22.
[4] Segura C, Palau À, Luque J, Costa-Jussà MR, Banchs RE. Chatbol, a chatbot for the Spanish "La Liga" In: Lecture Notes in Electrical Engineering. Springer; 2019. p. 319–30.
[5] Levin E, Levin A. Evaluation of spoken dialogue technology for real-time health data collection. J Med Internet Res. 2006;8(4):e30.
[6] Allen J, Ferguson G, Blaylock N, Byron D, Chambers N, Dzikovska M, Galescu L, Swift M. Chester: Towards a personal medication advisor. J Biomed Inform. 2006;39(5):500–13.
[7] Bazzi I. Modelling out-of-vocabulary words for robust speech recognition. Massachusetts Institute of Technology; 2002.
[8] Goldwater S, Jurafsky D, Manning CD. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Commun. 2010;52(3):181–200.
[9] Bickmore T, Trinh H, Asadi R, Olafsson S. Safety first: conversational agents for health care. In: Studies in Conversational UX Design. Springer; 2018. p. 33–57.
[10] Trujillo AL. Insulin analogs and pregnancy. Diabetes Spectr. 2007;20(2):94–101.
[11] Gong E, Baptista S, Russell A, Scuffham P, Riddell M, Speight J, Bird D, Williams E, Lotfaliany M, Oldenburg B. My Diabetes Coach, a Mobile App–Based Interactive Conversational Agent to Support Type 2 Diabetes Self-Management: Randomized Effectiveness-Implementation Trial. J Med Internet Res. 2020;22(11):e20322.
[12] Benzeghiba M, De Mori R, Deroo O, Dupont S, Erbes T, Jouvet D, Fissore L, Laface P, Mertins A, Ris C. Automatic speech recognition and speech variability: A review. Speech Commun. 2007;49(10–11):763–86.
[13] Myers C, Furqan A, Nebolsky J, Caro K, Zhu J. Patterns for how users overcome obstacles in voice user interfaces. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018. p. 1–7.
[14] Bohus D, Rudnicky A. Sorry and I Didn't Catch That!-An Investigation of Non-understanding Errors and Recovery Strategies. In: Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. 2005. p. 128–43.
[15] Bibault J-E, Chaix B, Nectoux P, Pienkowski A, Guillemasé A, Brouard B. Healthcare ex Machina: Are conversational agents ready for prime time in oncology? Clin Transl Radiat Oncol. 2019;16:55–9.
[16] Aberdeen J, Ferro L. Dialogue patterns and misunderstandings. In: ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems. 2003.
[17] Williams JD, Raux A, Henderson M. The dialog state tracking challenge series: A review. Dialogue & Discourse. 2016;7(3):4–33.
[18] Gamson K, Chia S, Jovanovic L. The safety and efficacy of insulin analogs in pregnancy. J Matern Neonatal Med. 2004;15(1):26–34.
[19] Higashinaka R, Funakoshi K, Araki M, Tsukahara H, Kobayashi Y, Mizukami M. Towards taxonomy of errors in chat-oriented dialogue systems. In: Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue. 2015. p. 87–95.
[20] Möller S, Engelbrecht K-P, Oulasvirta A. Analysis of communication failures for spoken dialogue systems. In: Eighth Annual Conference of the International Speech Communication Association. 2007.
[21] O'Halloran BM, Stone RB, Tumer IY. A failure modes and mechanisms naming taxonomy. In: 2012 Proceedings Annual Reliability and Maintainability Symposium. IEEE; 2012. p. 1–6.