L. Aunimo, "Enhancing Reliability and User Experience in Conversational Agents," *2023 World Symposium on Digital Intelligence for Systems and Machines (DISA*), Košice, Slovakia, 2023, pp. 16-18, doi: 10.1109/DISA59116.2023.10308922.

Enhancing Reliability and User Experience in Conversational Agents

Lili Aunimo

Haaga-Helia University of Applied Sciences, Ratapihantie 13, 00520 Helsinki, Finland

Abstract

Conversational agents are in place in a variety of domains and tasks such as sales and customer support services in business, student counselling in education and medical services in healthcare. There is abundant data available for modelling dialogs because online chat has been a popular way of communication between humans already for several decades. There are also innumerable other valuable digital resources that can be exploited when building a conversational agent, including pretrained large language models. We tested and evaluated several ways of preprocessing and modelling of chat dialogs in Finnish. As a result, we found out that the best accuracy is achieved using uncasing and spell-checking in the preprocessing phase and a BERT model pretrained with Finnish in the modelling phase. Despite the extensive use of conversational agents, there are still many open research questions. One example is the effect of the interaction style of the agent on user experience and emotions. Our initial study suggests that chatbots including small talk are less likely to elicit negative emotions, whereby emojis and emotional statements issued by chatbots do not play a significant role on the user's emotional responses. We also discuss how medical expert work may be partially automated and made more interesting as input for routine conversation is handled by a chatbot. Special attention is paid on the requirements for trustworthiness and reliability for conversational agents acting in different tasks and domains.

Keywords: conversational agent, language model, trustworthy AI, natural language processing, user experience, reliability

Introduction

Conversational agents (CAs) are widely used in many domains, such as healthcare, education, and retail, and for various kinds of tasks such as marketing, informing, counselling, and coaching. A CA may collaborate with a human e.g. when filling forms or screening for diseases. Electronic health records (EHR) are mostly filled in by medial doctors manually using a keyboard and a mouse. This work is away from the time they could spend interacting with their patients (Misrai et al., 2021). Voice-based CAs that interact with EHR databases would increase not only the productivity but also the quality of the work of the medical doctor. Due to the already currently wide usage of CAs and their future potential, it is important to study not only their technological aspects but also their reliability as well as the user experience and emotions triggered by them (Dobrowsky et al. 2019). Easy-to-use and emotionally sound CAs will contribute to productivity and wellbeing in the context of expert work automation (Aunimo et al., 2022).

Platforms and technologies employed to implement CAs are abundant (Adamopoulou, and Moussiades, 2020; Hussain et al., 2019). Modern CAs typically rely on large language models (LLMs) that have been finetuned for the specific task of CA. This study explores the use of a retrieval-based language model in medical chat. Retrieval-based chatbots can be considered more reliable than those based on generative models such as GPT (Generative Pretrained Transformers). Reliability is an important requirement in many professional contexts. Additional requirements on human oversight, safety, transparency, traceability, being non-discriminant and friendliness for the environment are posed by the forthcoming AI Act that categorizes AI systems into three categories (unacceptable risk applications, high-risk applications and limited or low-risk applications) based on the risk they pose to users (Veale and Borgesius, 2021). There are domains, such as healthcare and education, where AI systems are typically classified under the "high risk" category, meaning that they will have to undergo an assessment ensuring that the requirements are fulfilled before entering the market and throughout their lifecycle. Besides the upcoming AI Act, also the ethical guidelines for trustworthy AI pose requirements for CAs. Fulfilling the above requirements needs novel methods and techniques as the challenges are not by any means completely solved by existing techniques such as ChatGPT based CAs.

Conversational Agents: Definitions and Trustworthy AI

CAs have been extensively studied (Allouch et al., 2021) ever since there has been interest in intelligent behaviour by machines. Natural language conversation by a machine represents a classical field of AI as the definition of an intelligent machine has been defined as a system that can pass the Touring test (Touring, 1950). The term CA may have been used relatively scarcely, but in a broader sense CAs such as ELIZA (Weizenbaum, 1966) and ALICE (Wallace, 2009) are among the classical examples of major developments in artificial intelligence research. Research on question-answering (QA) systems represents a specific type of CAs, namely two-turn conversation. This line of research has been going on already since 1960s (Mishra and Jain, 2016). The only QA system for the Finnish language that has been evaluated using a common dataset for benchmarking was presented at the QA@CLEF Cross-Language Evaluation Forum (Vallin et al., 2006) by Aunimo et al. (2004).

The term CA has been used in several different contexts and with different meanings. In this paper, we mean by a CA a non-embodied agent that also uses natural language. It can understand utterances other than those given as input to it when building the agent, thus demonstrating a form of intelligence. Figure 1 shows a taxonomy of agents that employ natural language in a form or another. This type of agents may also be called dialog systems. Dialog systems are divided into two categories: CAs and interactive voice response systems (IVR) (Allouch et al., 2021). An IVR does not understand natural language or generate it. It just handles predefined input and output. A virtual call-center receptionist is an example of an IVR. It handles predefined commands such as: "If you would like to have service in English, press 1, otherwise press 2". A CA in turn is a dialogue system that can also understand and generate natural language content, using text, voice, or hand gestures, such as sign language (Allouch et al., 2021). CAS are classified into text-based agents, voice-based agents and embodied agents. Embodied agents may either be graphically or physically embodied (Allouch et al., 2021). This paper deals with all types of CAs except the physically

embodied ones. Thus, also multimodal CAs are in scope. Multimodal CAs communicate with text, speech, signs and gestures.

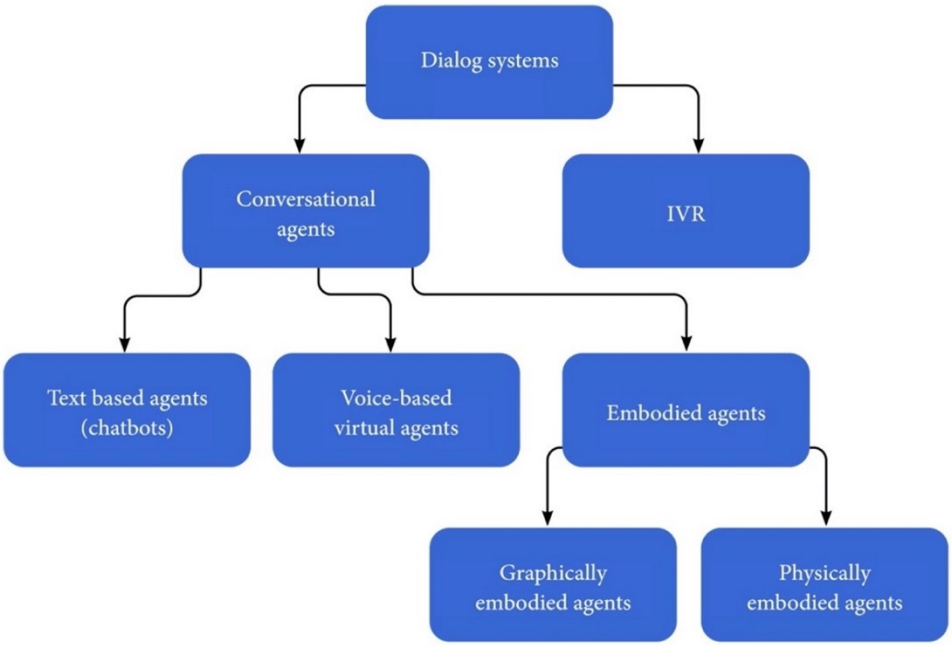

Figure 1: A typology of dialogue systems. Source: Allouch, Azaria and Azoulay (2021). CAs are one type of dialogue systems.

Many professional contexts present strict requirements for the trustworthiness of CAs. All the seven requirements for trustworthy AI defined by the High-Level Expert Group on Artificial Intelligence set by the European Commission in their "Ethics Guidelines for Trustworthy AI" (High-Level Expert Group on AI, 2019) should in many cases be put into practice. The seven key requirements are: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability (Ethics, 2019). These requirements are not easy to fulfill. For example, achieving both privacy and transparency when building a model based on real data and machine-learning methods is challenging. There are also domain-specific challenges, such as many patient groups being vulnerable and thus the need to pay special attention to ensure human agency and non-discrimination, is present in most situations.

Experiments on a Retrieval-based Conversational Agent for Finnish

We performed experiments for a retrieval-based CA by taking Finnish multi-turn dialogue data and building a CA for partly automating the work of a medical doctor. The data set consisted of 29602 dialogues between doctor and a patient. Based on this data, be built several models using machine learning methods and various preprocessing techniques. The goal was to compare the performance of these techniques. The ultimate task was to predict the next utterance of the medical doctor in a dialogue between him and a patient. The experiments and results are discussed thoroughly in the paper Kauttonen and Aunimo (2019). The best performing combination of preprocessing and modelling was spell-checking without lemmatization and FinBERT as the model. This combination reached an accuracy of 92% in the multi-turn next utterance prediction task, meaning that in 92 cases out of 100,

the correct response was ranked as the first one among the ten utterances predicted by the model (1 of 10 Recall@1 accuracy).

The fact that FinBERT provided better results than the multilingual BERT is in line with the results of other researchers (Canete et al., 2023). Language-specific BERT has been found to perform better than the multilingual BERT in several NLP tasks for French, Spanish, Dutch and Portuguese, among others.

Our results for the experiments regarding preprocessing are not very definitive and more thorough experiments are needed to make stronger conclusions. In our study, the best results were obtained when the tokens were not lemmatized but left as they were. Adding spell-checking to the raw tokens did not provide any noteworthy difference in the performance. Lemmatization typically reduces the size of the vocabulary in morphologically rich languages (Kanerva et al., 2018). However, in the case of the medical chat the vocabulary size was relatively small when compared with the vocabulary size of the ask a librarian task (Kauttonen and Aunimo, 2019). We can observe that the difference in performance between the raw and lemmatized is smaller in the ask a librarian data set. The reason for the good performance on raw data may be that BERT uses word pieces to handle out-of-vocabulary words (Devlin et al., 2019).

A retrieval-based CA may be more trustworthy than a generative one because the answers it gives have once been approved by a medical doctor and they exist in the training corpus. CAs that use generative models are more flexible when unseen dialog turns appear. However, also they are limited by their training data. One of the main challenges with CAs based on generative pretrained transformers (also called large language models) such as those used in ChatGPT by Open AI [1] is that they may produce erroneous answers (Ji et al., 2023) or safe answers with very little information content. However, CAs using generative models may employ methods for ensuring that there is some information content in the utterances (Mou et al., 2016). The correctness of dialogue output may also be judged by using a knowledge base or performing fact checking after retrieving the answer (Gupta et al., 2021). Additionally, fact-checking may be performed using a trusted source such as Wikipedia (Kim et al., 2021). All these methods could be experimented with to ensure the reliability of information produced by a CA.

Interaction Style of the Conversational Agent and Emotional Response

We studied the interaction between a CA and an information seeker. Altogether 78 informants were given an information seeking task and then randomly assigned one of the four different CA variants. At the first level, a neutral CA which generates only factual information without any expression of emotion, and a positively valenced CA which enriches communication with short emotional statements and emojis was implemented. At the second level, either a small talk module was added or not. Data were collected through a 2x2 between-subjects factorial design (neutral versus emotional; small talk vs. no small talk). 78 participants were randomly assigned to one of the four experimental groups. Participants

[1] https://openai.com/blog/chatgpt

were asked to perform the information seeking task through CA interaction. During the interaction, their facial expressions were recorded via a webcam. The videos were then analyzed using the AFFDEX algorithm, which assigns 34 facial expressions to 7 basic emotions (anger, sadness, disgust, joy, surprise, fear, contempt). The algorithm builds on Emotional Facial Action Coding System (EMFACS) mappings developed by Ekman et al. (1994). The experiment was implemented using the iMotions software[2].
Although no group differences are confirmed at the level of individual emotions, the results suggest that CAs including small talk are least likely to elicit negative emotions, but emojis and emotional statements did not play a significant role. The results suggest that the interaction style of a CA does affect the emotional response of the user. This is a first step in demonstrating a way to capture and analyze emotional reactions of CA users by employing facial expression analysis. Even though results should be interpreted with caution due to the sample size, they provide initial indications in terms of CA interaction design.

Conclusions

This paper presented two experiments involving CAs. The first one discussed a retrieval-based CA in a medical context, as an assistant for the medical doctor working in telemedicine and in the Finnish language. In this context reliability of the information and its trustworthiness as defined by the Expert Group on AI Ethics are of outmost importance. The user experience of a medical doctor is also important. Techniques for creating trustworthy and reliable CAs exist, but the challenges are far from being solved – even seen the recent developments in CAs based on LLMs.

The second experiment concentrated on user experience and examined the interaction between the informant and the CA from the point of view of emotions expressed by the informant. This is an important line of study as the time and effort spent interacting with CAs grows all the time. The initial results show that people do express emotions while interacting with CAs. This is not evident, as facial expressions typically are used for communicating with other humans. We found some phases in the interaction that produced negative emotions in many of the informants. We also found out that the CA variant with small talk introduced fewer negative emotions than the one with no small talk module. This piece of information is useful for those designing the interactions for CAs. This is also interesting from the methodological point of view as emotion detection has not been widely used in CA research.

References

Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. In IFIP international conference on artificial intelligence applications and innovations (pp. 373-383). Springer, Cham.

Allouch, M., Azaria, A., & Azoulay, R. (2021). Conversational agents: Goals, technologies, vision and challenges. Sensors, 21(24), 8448.

[2] https://imotions.com/

Aunimo, L., Kauttonen, J., & Alamäki, A. (2022). Expert Work Automation in Healthcare: the Case of a Retrieval-Based Medical Chatbot.

Aunimo, L., Kuuskoski, R., & Makkonen, J. (2004). Cross-Language Question Answering at the University of Helsinki. In CLEF (Working Notes).

Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2023). Spanish pre-trained bert model and evaluation data. arXiv preprint arXiv:2308.02976.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423

Dobrowsky, D., Aunimo, L., Janous, G., Pezenka, I., & Weber, T. (2020). The Influence of Interactional Style on Affective Acceptance in Human-Chatbot Interaction – A Literature Review. *AINL: Artificial Intelligence and Natural Language Conference*. Workshop on Human-AI Interaction. 7.-9.10.2020, online. http://urn.fi/URN:NBN:fi-fe2021101451016

Ekman, P., Irwin, W., & Rosenberg, E. (1994). The emotional facial action coding system (EMFACS). *London, UK*.

Gupta, A., Varun, Y., Das, P., Muttineni, N., Srivastava, P., Zafar, H., ... & Nath, S. (2021). TruthBot: An Automated Conversational Tool for Intent Learning, Curated Information Presenting, and Fake News Alerting. arXiv preprint arXiv:2102.00509.

High-Level Expert Group on AI (2019). *Ethics guidelines for trustworthy AI* (Report). European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hussain, S., Ameri Sianaki, O., & Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33 (pp. 946-956). Springer International Publishing.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., & Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics.

Kauttonen, J., & Aunimo, L. (2020). Dialog modelling experiments with finnish one-to-one chat data. In Conference on Artificial Intelligence and Natural Language (pp. 34-53). Cham: Springer International Publishing.

Kim, B., Kim, H., Hong, S., & Kim, G. (2021). How Robust are Fact Checking Systems on Colloquial Claims?. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1535-1548).

Mishra, A., & Jain, S. K. (2016). A survey on question answering systems with classification. Journal of King Saud University-Computer and Information Sciences, 28(3), 345-361.

Misrai, V., Pon, D., & Charbonneau, H. (2021). While the Chatbot's Away, the Mice Will Play. Frontiers in Digital Health, 3, 617013.

Mou, L., Song, Y., Yan, R., Li, G., Zhang, L., & Jin, Z. (2016). Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. arXiv preprint arXiv:1607.00970.

Turing, A. (1950) Computing machinery and intelligence. Mind 59, 433–460.

Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4), 97-112.

Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., ... & Sutcliffe, R. (2006). Overview of the CLEF 2005 multilingual question answering track. In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers 6 (pp. 307-331). Springer Berlin Heidelberg.

Wallace, R. S. (2009). The anatomy of ALICE (pp. 181-210). Springer Netherlands.

Weizenbaum, J. (1966) ELIZA: a computer program for the study of natural language communication between men and machines. Commun. ACM 9, 36–45.