Systematic review of training methods for conversational systems: the potential of datasets validated with user experience

Carolina Abrantes*, Juliana Camargo**, José Nunes***, Maria João Antunes****, Óscar Mealha*****, Luís Nóbrega******

* GoTelecom, S.A., Aveiro (cabrantes@broadvoice.com)

** Department of Communication and Art/DigiMedia, Aveiro University (julianacamargo@ua.pt)

*** Department of Communication and Art/DigiMedia, Aveiro University (jnunes@ua.pt)

**** Department of Communication and Art/DigiMedia, Aveiro University (mariajoao@ua.pt)

***** Department of Communication and Art/DigiMedia, Aveiro University (oem@ua.pt)

****** GoTelecom, S.A., Aveiro (lnobrega@broadvoice.com)

Abstract

The maturation of artificial intelligence technologies, such as Machine Learning algorithms, Natural Language Processing (NLP), Automatic Speech Recognition (ASR), and Natural Language Generation, is transforming the way users interact with technology. With the increasing prevalence of voice interactions, it is crucial to understand the process of training conversational agents. This systematic review examines how human data is collected for training these agents, with a specific focus on datasets obtained directly from human participation in real-life contexts of need and use. The study follows the PRISMA guidelines, and searches were conducted in Scopus, Web of Science, and ProQuest databases in English, covering the period from 2005 to 2020, to provide a comprehensive overview of published practices until July 2020. A total of 22 papers were included in this review from the search iterations. The primary findings of these papers indicate a common use of learning from demonstration/observation and crowdsourcing methods in system training and dataset cataloguing. Additionally, techniques such as handwriting and sentence labelling, as well as Wizard-of-Oz based studies, were employed in the research.

Keywords: Artificial Intelligence; conversational systems; datasets; human data.

Introduction

Conversing with appliances, machines, and computers has become an increasingly common practice (Silva et al., 2020). This is possible due to advances in Natural Language Processing (NLP), a field of computing that enables the understanding of voice commands, allowing users to interact using natural language phrases (Rani, Bakthakumar, Kumaar, Kumaar, & Kumar, 2017).

The popularity of this feature has grown due to the convenience it offers, as users no longer need to rely on manual input for giving commands or scheduling tasks. Intelligent personal assistants (IPAs), such as Alexa and Siri, are being used more frequently by diverse audiences. Speech has even become the primary form of interaction with home devices like Amazon Echo, Google Home, and Apple HomePod (Clark et al., 2019). This new mode of interaction is a consequence of advances in Artificial Intelligence (AI).

To understand and respond to user commands, AI-driven machines require databases that are continuously enriched with new information generated from the use of the AI system itself (West, 2019). For instance, streaming platform Netflix suggests content based on a user's viewing history.

Regarding speech-based resources, AI employs NLP to comprehend human-spoken commands and simulate appropriate responses. Chatbots, known as conversational agents, are an example of this approach. These systems are designed using characteristics of AI, Machine Learning, and NLP to receive user input and provide suitable responses, simulating a human conversation (Medeiros & Bosse, 2018).

Since each system is used in a different context, its adaptability requires specific data and a thorough training process for the Natural Language Understanding (NLU) component, often involving human collaboration (Santos, Abreu, & Almeida, 2019).

This systematic literature review contributes to the existing knowledge by understanding how human data is collected for training conversational agents. To define the search keywords, an open exploratory exercise was conducted across various databases to identify the most frequently used terms in scientific publications related to this subject.

Following the selection of relevant terms, this review was conducted in accordance with the PRISMA guidelines, addressing both quantitative and qualitative aspects of a set of 22 articles identified in the SCOPUS, ProQuest, and Web of Science databases. The paper is divided into four parts. Firstly, the methodological procedures are described in detail. Secondly, the search results are presented.

The third part highlights the main findings of this review. Finally, a discussion of the results and the conclusions summarize the key contributions of this review. The corpus of scientific papers used in this work revealed that crowdsourcing was the most common method integrating user experience for training systems that enable voice interactions, highlighting its ability to adapt to specific contexts and situations.

Method

The systematic literature review's primary objective was to identify how human data is collected to train conversational systems. For this, we chose to select and perform the meta-analysis of the scientific papers based on the PRISMA search guidelines as proposed by Moher et al. (2009). The search was centered on three databases: Scopus, Web of Science and ProQuest (the language chosen was English).

To obtain results consistent with the current reality, we chose to define the last 15 years (2005-2020) as our search period. On the other hand, there were no limitations of either, scientific area or type of document, all results were considered.

The groups of terms, listed in Table 1, were based on the existing literature, an outcome of the initial exploratory exercise, as mentioned in the previous section. The selected words were introduced in the search equation combined with Boolean operators (e.g., "OR" and AND"). All searches were limited to the title, summary, and keywords of the articles.

Table 1 – Keywords defined for the systematic literature review

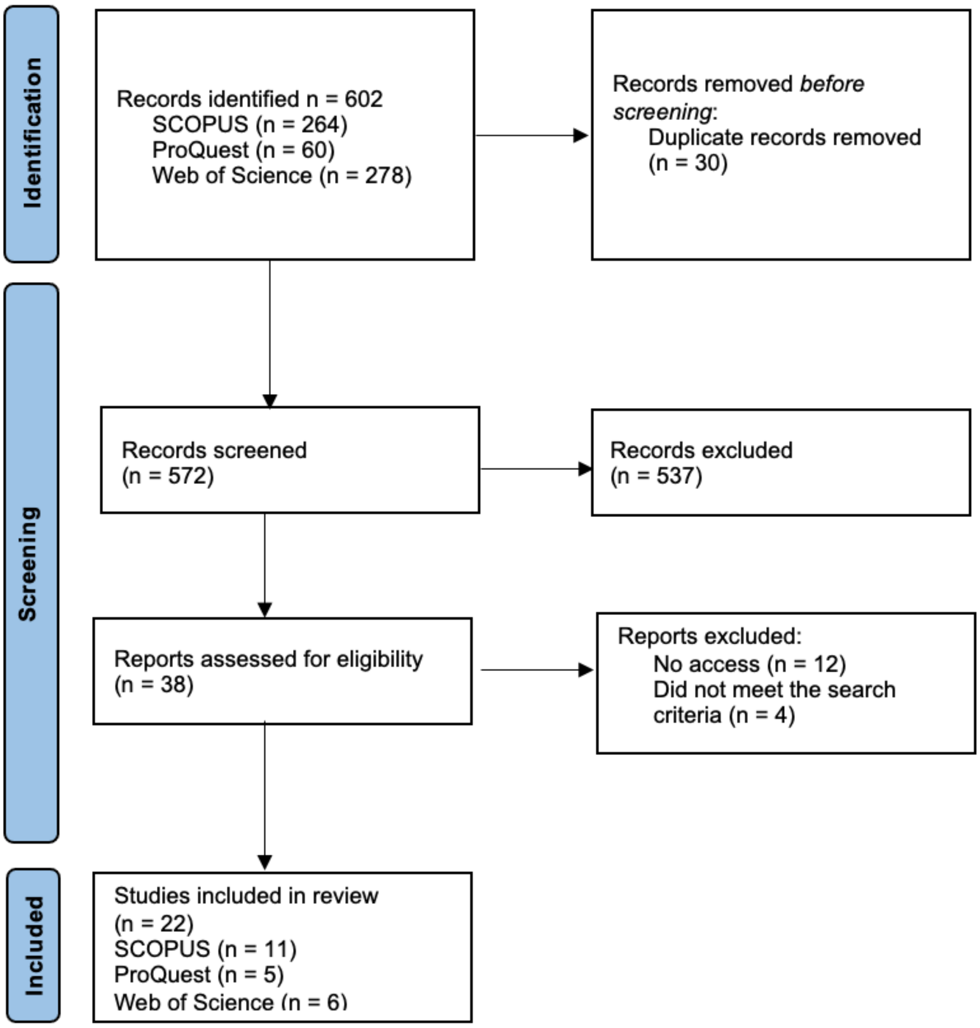
Group 1 (agents or bots): Voice, conversational, utterance, bot, NLP, design, training, building, techniques, methods, tools, intent, entity. Group 2 (human support): Data setup, learning process, demonstration, neural network, dataflow model, intelligent systems, supervised learning, reinforcement learning, conversation, voice processing, speech recognition, dialogue system, user, human, behaviour. These changes ensure that the nouns and verbs are in the appropriate form (e.g., "building" instead of "build", "techniques" instead of "technique", etc.).

The following criteria were considered to select papers for this review:

i) the reported method uses empirical data resulting from the observation or measurement of human behavior;
ii) the reported work has a system training objective;
iii) the paper clearly states how the system is trained;
iv) the paper clearly states how human data is used to train a conversational system.

A total of 602 results were identified in three databases, as depicted in Figure 1. After applying the selection criteria described in the previous section and eliminating 30 duplicate papers, 572 papers remained. Among these, 537 were excluded based on the analysis of the title and abstract.
Subsequently, of the remaining 35 papers, 12 were removed due to lack of access to the full version of the document. After reading the remaining 23 papers, 4 were excluded for not meeting the research criteria. Finally, during the reading of the papers, 3 additional papers were identified and considered relevant to the research. These references were included in the systematic review process and went through the complete flow, starting from step 1.

Figure 1 - Search flow in the SCOPUS, ProQuest and Web of Science databases



Source: authors, adapted from Prisma Method by Moher et al. (2009)

The following section explains how the included studies depicted in Figure 1 (22 accepted papers) were obtained from the search process. The main themes are clustered within the context of the accepted papers.

Results

The results are organized into two sections to provide a better explanation of both search iterations conducted on the three databases. Both iterations yielded 11 papers each, resulting in a total of 22 papers as previously indicated in Figure 1. The key findings are categorized and presented in tables, indicating the number of unique papers in which they were found.

First search iteration: Training methods for conversational systems

To investigate the current practices employed in training conversational agents, two queries were conducted in each of the databases. Initially, keywords from the first group were combined, and then the terms "intent" and "entity" were added as they are commonly used in this context to refer to user requests made to the system (Gnanaguru, 2020). After inspecting the abstracts, a total of 15 articles were selected.

Among these, 5 were excluded due to limited access to the full papers, 2 were excluded after a thorough review of their full content, and 3 were added from the references of the accepted papers. Consequently, the first search process yielded 11 selected papers for the meta-analysis.

The main themes extracted from these papers were categorized based on their purpose (i.e., which system was being trained or the main reported goal), training method, and system analysis method (not a systematic review objective). The clusters resulting from the analysis of the first iteration's 11 papers are presented in Table 2, along with the corresponding number of unique papers in which each cluster was found. It's worth noting that some papers may fall into multiple categories simultaneously.

Table 2 – Key findings from the 1st systematic review iteration, clusters, and related number of unique papers

Purpose (system/paper)	
Chatbot	3
Method to generate initial discourse data	3
Method to identify intents and entities	3
Conversational system	2
Intent classification method	1
Training/research methods	
Study of current practices	3
Crowdsource	3
Artificial/deep neural network	2
Coded	1
Labelled transcribed dialogues	1
Dataset of example utterances	1
Analysis method	
Wizard-of-Oz	1

The accepted papers primarily focused on the following research areas: chatbot training (3 papers), methods for generating a dataset of initial bot utterances (3 papers), and techniques for identifying intents and entities (3 papers). Additionally, there were 2 papers that addressed the training of conversational systems without specifying the exact scope, while 1 paper presented a method for intent classification based on user utterances.

Relating to methods, 3 articles reported on current practices, 3 used crowdsourcing for bot training, 2 employed neural networks, 1 involved manual coding of the bot (without a continuous learning mechanism), 1 relied on manual labeling of transcribed dialogues, and 1 paper mentioned having a dataset of example utterances.

Not being a primary research objective, it is noteworthy to highlight that one study reported using the Wizard-of-Oz method to analyze their final conversational agent. This method involves participants interacting with an interface, unaware that the system feedback is actually being provided by an unseen human rather than the computing system.

Second search iteration: Methods for system training based on human empirical data

The second iteration of the systematic review aimed to examine how human data is collected and used in other areas of system training. To achieve this, two queries were employed. The first query included keywords such as data setup, learning process, demonstration, neural networks, dataflow model, intelligent systems, supervised learning, and reinforcement learning, in combination with terms like conversation, voice processing, speech recognition, and dialogue system to gather relevant studies related to communication. For the second query, the communication-related keywords were replaced with user, human, and behavior to gain a broader perspective.

The analysis of abstracts resulted in the acceptance of 20 papers. However, 7 of these papers were unavailable despite requests and were subsequently excluded. After conducting a full review of the remaining articles, 2 additional papers were found but did not meet the inclusion criteria and were excluded as well. No papers were added from the references. In total, 11 additional papers were accepted for the review process.

The clustering of themes in this iteration followed the same criteria as the first systematic review iteration, focusing on the main purpose or system being studied, the stated training method(s), and, if mentioned, the analysis method of the respective system. All these clusters are detailed in Table 3.

Table 3 **-** Key findings from the 2nd systematic review iteration, clusters and related number of unique papers

Purpose (system)	
Humanoid robot	2
Human interaction simulator	2
Assistive system	2
In-car multimodal system	1
NPC bot (gaming)	1
Recommender system	1
Chatbot	1
Training methods	
Demonstration from an expert (behavior and/or speech)	7
Speech capture (utterances)	2
Example images	1
Sample dialogues	1

Interaction logs (interaction with interface)	1
Coded, autonomous	1
Analysis method	
Turing test	1

In terms of system purpose, 2 papers are focused on humanoid robots, 2 papers discuss human interaction simulators, and 2 papers are related to assistive systems. Additionally, there is 1 paper each for the following themes: in-car systems, NPC bot, recommender system, and chatbot.

Regarding training methods, 7 studies report the use of learning from demonstration methods, which include both physical behavior and speech. Furthermore, 2 studies specifically address speech (utterance) capture. Each of the following themes is mentioned in 1 paper: example images, sample dialogues, interaction logs obtained from user interface interaction, and a coded mechanism with autonomous features. One of the studies also refers to Turin tests as an analysis method.

Findings and meta-analysis

This section aims to provide a comprehensive expansion of the systematic review findings by discussing the key trends that have been identified and briefly summarized. The first subsection introduces a state-of-the-art overview of systems training using human empirical data, providing a broader perspective. The second subsection delves into more specific methods for setting up conversational systems based on user experience.

System training based on human empirical data

Among the various research contexts, the most commonly observed method of system training based on human data is learning from demonstration/observation, as evidenced by seven different documents (refer to Table 3). Less frequently mentioned methods include the use of interaction logs (Z. Li, Kiseleva, De Rijke, & Grotov, 2017) and autonomous data collection with network-shared knowledge (Ahrndt, Lützenberger, & Prochnow, 2016).

Learning from demonstration appears to be particularly prevalent in the training of humanoid robots. These robots, equipped with deep reinforced learning capabilities, are capable of speaking, listening, gesturing, and learning. The training process involves utilizing example images, labeled videos, dialogues, or physical demonstrations performed by humans (Cuayáhuitl, 2020; G. Li, He, Gomez, & Nakamura, 2018; Moro, Nejat, & Mihailidis, 2018).

This training method is also applied in digital robots, such as those used in gaming environments (Lee, Luo, Zambetta, & Li, 2014), assistive agents (Ahrndt et al., 2016; Diaz, Girgis, Fevens, & Cooperstock, 2017; Fenza, Orciuoli, & Sampson, 2017), and human communication simulators (Novikova, Lemon, & Rieser, 2016; Vanzo, Part, Yu, Nardi, & Lemon, 2018).

Neural networks play a crucial role in capturing the ideal behavior demonstrated by humans (Diaz et al., 2017). As mentioned by G. Li et al. (2018), when agents operate in human environments, which are prevalent in real-world applications, the ability to interact successfully becomes a key factor for system success.

An interesting method in the context of learning dialogue policies is applied by Doering, Kanda, & Ishiguro (2019). In their work, a physical scenario is configured to collect human-human interactions, which are then used to simulate speech interactions in purchasing scenarios.

Conversational systems setup

Focusing on the main objectives of chatbots, Ayanouz, Abdelhakim, & Benhmed (2020) highlight several limitations inherent to these systems. They mention the rigidity of rules, which hinders their ability to respond effectively to unforeseen scenarios, as well as the lack of contextual understanding, which leads to various conversation issues. Huang, Lasecki, & Bigham (2015) further emphasize that existing dialogue systems have limited scope, struggle with the complexity of natural language, and are expensive to develop.

To address some of these challenges, conversational systems heavily rely on a comprehensive knowledge base. Bapat, Kucherbaev, & Bozzon (2018) emphasize that the user experience of chatbots is highly dependent on the performance of the Natural Language Understanding (NLU) model, which in turn heavily relies on the initial dataset used for training. These datasets consist of user utterances, including labeled phrases, paraphrases, and their corresponding intents, along with grammatical variations and prompts (Wang, Bohus, Kamar, & Horvitz, 2012).

Methodological approach

Two main approaches related to the collection of user experience data are outlined in this review: crowdsourcing and handwriting and sentence labeling, with some cases integrating both tactics. Additionally, the Wizard-of-Oz approach is used for data collection and system analysis, although it can be costly.

Crowdsourcing

In crowdsourcing approaches, participants are provided with an initial utterance and asked to paraphrase it into new expressions and variations. Wang et al. (2012) explored different methods in crowdsourcing data acquisition, including providing an initial sentence, scenarios (storytelling), or list-based descriptions.

The diversity of user requests should be covered in the initial training data. Crowdsourcing approaches have the potential to enable conversational systems to continuously learn and improve over time. Huang et al. (2015) introduced a framework that involves gathering information from web APIs through crowd validation and translation of API results into natural language sentences.

Handwriting and sentence labeling

Expert-built conversational templates have limitations as they address only a small subset of ways to refer or ask for information. In this method, domain experts generate hand-crafted, domain-specific grammars and placeholder dialogue templates.

These systems are not scalable, highly domain/context-specific, and costly. Gupta et al. (2006) propose a scalable system that uses transcription and analysis of user speech data to create a classifier model for understanding callers' intents and named entities. The classifier model integrates hand-crafted classification rules with rules learned from data.

Wizard-of-Oz

The Wizard-of-Oz approach involves collecting speech data and analyzing the final system, but it can be expensive. It allows researchers to understand user interactions and fine-tune prompts, gather intents, utterances, and entities.

Ayanouz et al. (2020) suggest that end-to-end neural networks are replacing manual methods, proposing an API with an initial dataset and a delexicalization technique for incremental extension. Leggeri et al. (2018) propose a method based on sentiment analysis to dynamically increase and adapt the communication dataset in real time.

Discussion

Training an AI conversational system requires user utterances, intents, and entities. These parameters are dependent on user dialogue, emphasizing the importance of nurturing the system with human data gathered from natural interaction contexts and situations.

Employing human-centered, user experience research techniques could be promising for gathering data in smaller and context-specific interaction scenarios. Interviews and workshops with experts and local clients can help understand communicational needs and experiences. Understanding the why and the specific social or cultural context of conversational instances is crucial. Systems trained with person-centered dialogue methods may be better accepted by potential end-users.

Conclusion

The systematic literature review presented in this article provides compelling evidence for the recurrent use and positive outcomes of training systems based on human empirical data. While the review is limited to specific databases, a total of 22 articles met the inclusion criteria outlined in this study, shedding light on the trends in incorporating human-based data collection methods for training AI systems.

Among the identified methods, learning from demonstration and crowdsourcing emerged as the most common approaches for training intelligent systems with human data. The review identified 7 studies utilizing learning from demonstration and 3 studies employing crowdsourcing as effective strategies for system training.

By leveraging these methods, AI systems can acquire a deeper understanding of how real people communicate, enabling them to better comprehend user requests. This finding underscores the significant potential of the reported methods in training AI-based systems, particularly in smaller or specialized contexts.

Acknowledgements

This work is supported by the European Regional Development Fund (ERDF), through the Competitiveness and Internationalization Operational Programme (COMPETE 2020) of the Portugal 2020 framework [Project

CallBoB with Nr. 038501 (POCI-01-0247-FEDER-038501)] and is also financially supported by national funds through FCT – Foundation for Science and Technology, I.P., under the project UIDB/05460/2020.

Bibliographical references

Ahrndt, S., Lützenberger, M., & Prochnow, S. M. (2016). Using personality models as prior knowledge to accelerate learning about stress-coping preferences (demonstration). In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*.

Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020). A Smart Chatbot Architecture based NLP and Machine Learning for Health Care Assistance. In *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3386723.3387897

Bapat, R., Kucherbaev, P., & Bozzon, A. (2018). Effective crowdsourced generation of training data for chatbots natural language understanding. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-91662-0_8

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., … R Cowan, B. (2019). The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*. https://doi.org/10.1093/iwc/iwz016

Cuayáhuitl, H. (2020). A data-efficient deep learning approach for deployable multimodal social robots. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.09.104

de Barcelos Silva, A., Gomes, M. M., da Costa, C. A., da Rosa Righi, R., Barbosa, J. L. V., Pessin, G., … Federizzi, G. (2020). Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2020.113193

Diaz, M., Girgis, R., Fevens, T., & Cooperstock, J. (2017). To Veer or Not to Veer: Learning from Experts How to Stay Within the Crosswalk. In *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*. https://doi.org/10.1109/ICCVW.2017.174

Doering, M., Kanda, T., & Ishiguro, H. (2019). Neural-network-based Memory for a Social Robot. *ACM Transactions on Human-Robot Interaction*. https://doi.org/10.1145/3338810

Fenza, G., Orciuoli, F., & Sampson, D. G. (2017). Building Adaptive Tutoring Model Using Artificial Neural Networks and Reinforcement Learning. In *Proceedings - IEEE 17th International Conference on Advanced Learning Technologies, ICALT 2017*. https://doi.org/10.1109/ICALT.2017.124

Gnanaguru, G. (2020). Building a ChatBot Using Emotional Cognitive Conversational Agent Architecture (ECCAA). In *Lecture Notes in Electrical Engineering*. https://doi.org/10.1007/978-981-15-1420-3_187

Gupta, N., Tur, G., Hakkani-Tür, D., Bangalore, S., Riccardi, G., & Gilbert, M. (2006). The AT&T spoken language understanding system. In *IEEE Transactions on Audio, Speech and Language Processing*.

https://doi.org/10.1109/TSA.2005.854085

Huang, T. K., Lasecki, W. S., & Bigham, P. (2015). Guardian : A Crowd-Powered Spoken Dialog System for Web APIs. *AAAI Conference HCOMP*.

Lee, G., Luo, M., Zambetta, F., & Li, X. (2014). Learning a Super Mario controller from examples of human play. In *Proceedings of the 2014 IEEE Congress on Evolutionary Computation, CEC 2014*. https://doi.org/10.1109/CEC.2014.6900246

Leggeri, S., Esposito, A., & Iocchi, L. (2018). Task-oriented conversational agent self-learning based on sentiment analysis. In *CEUR Workshop Proceedings*.

Li, G., He, B., Gomez, R., & Nakamura, K. (2018). Interactive Reinforcement Learning from Demonstration and Human Evaluative Feedback. In *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*. https://doi.org/10.1109/ROMAN.2018.8525837

Li, Z., Kiseleva, J., De Rijke, M., & Grotov, A. (2017). Towards learning reward functions from user interactions. In *ICTIR 2017 - Proceedings of the 2017 ACM SIGIR International Conference on the Theory of Information Retrieval*. https://doi.org/10.1145/3121050.3121098

Medeiros, L., & Bosse, T. (2018). An intelligent conversational bot to help users cope with stress online. In *Belgian/Netherlands Artificial Intelligence Conference*.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.1000097

Moro, C., Nejat, G., & Mihailidis, A. (2018). Learning and Personalizing Socially Assistive Robot Behaviors to Aid with Activities of Daily Living. *ACM Transactions on Human-Robot Interaction*. https://doi.org/10.1145/3277903

Novikova, J., Lemon, O., & Rieser, V. (2016). Crowd-sourcing NLG Data: Pictures Elicit Better Data. https://doi.org/10.18653/v1/w16-6644

Rani, P. J., Bakthakumar, J., Kumaar, B. P., Kumaar, U. P., & Kumar, S. (2017). Voice controlled home automation system using natural language processing (NLP) and internet of things (IoT). In *ICONSTEM 2017 - Proceedings: 3rd IEEE International Conference on Science Technology, Engineering and Management*. https://doi.org/10.1109/ICONSTEM.2017.8261311

Romero, M., Casadevante, C., & Montoro, H. (2020). CÓMO CONSTRUIR UN PSICÓLOGO-CHATBOT. *Papeles Del Psicólogo - Psychologist Papers*. https://doi.org/10.23923/pap.psicol2020.2920

Santos, R., Abreu, J. F. De, & Almeida, P. (2019). Training Natural Language Understanding for the TV Context : a Visual Stimuli Approach for the Elicitation Process The Use of Visual Stimuli for NLU Training. *Applications and Usability of Interactive TV*, *2*, 1–14. https://doi.org/10.1007/978-3-030-56574-9_6

Shin, C. U., & Cha, J. W. (2019). End-to-end task dependent recurrent entity network for goal-oriented dialog learning. *Computer Speech and Language*. https://doi.org/10.1016/j.csl.2018.06.004

Tran, O. T., & Luong, T. C. (2020). Understanding what the users say in chatbots: A case study for the Vietnamese language. *Engineering Applications of Artificial Intelligence*. https://doi.org/10.1016/j.engappai.2019.103322

Vanzo, A., Part, J. L., Yu, Y., Nardi, D., & Lemon, O. (2018). Incrementally learning semantic attributes through dialogue interaction: Robotics track. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Wang, W. Y., Bohus, D., Kamar, E., & Horvitz, E. (2012). Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *2012 IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings*. https://doi.org/10.1109/SLT.2012.6424200

West, E. (2019). Amazon: Surveillance as a service. *Surveillance and Society*. https://doi.org/10.24908/ss.v17i1/2.13008

Yaghoub-Zadeh-Fard, M. A., Benatallah, B., & Zamanirad, S. (2020). Automatic canonical utterance generation for task-oriented bots from API specifications. In *Advances in Database Technology - EDBT*. https://doi.org/10.5441/002/edbt.2020.02

Yan, R. (2018). "Chitty-chitty-chat bot": Deep learning for conversational AI. In *IJCAI International Joint Conference on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2018/778