DOI: 10.1609/aaai.12007

CONVERSATIONAL AI

Considerations on creating conversational agents for multiple environments and users

Javier Cebrián[1] | **Ramón Martínez**[1] | **Natalia Rodríguez**[1] | **Luis Fernando D'Haro**[2]

[1] Saturno Labs, Villaviciosa de Odón, Madrid, Spain

[2] ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain

Correspondence
Javier Cebrián, Saturno Labs, Villaviciosa de Odón 28670 Madrid, Spain.
Email: javier@saturnolabs.com

Luis Fernando D'Haro, ETSI de Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain.
Email: luisfernando.dharo@upm.es

Abstract

Advances in artificial intelligence algorithms and expansion of straightforward cloud-based platforms have enabled the adoption of conversational assistants by both, medium and large companies, to facilitate interaction between clients and employees. The interactions are possible through the use of ubiquitous devices (e.g., Amazon Echo, Apple HomePod, Google Nest), virtual assistants (e.g., Apple Siri, Google Assistant, Samsung Bixby, or Microsoft Cortana), chat windows on the corporate website, or social network applications (e.g. Facebook Messenger, Telegram, Slack, WeChat).

Creating a useful, personalized conversational agent that is also robust and popular is nonetheless challenging work. It requires picking the right algorithm, framework, and/or communication channel, but perhaps more importantly, consideration of the specific task, user needs, environment, available training data, budget, and a thoughtful design.

In this paper, we will consider the elements necessary to create a conversational agent for different types of users, environments, and tasks. The elements will account for the limited amount of data available for specific tasks within a company and for non-English languages. We are confident that we can provide a useful resource for the new practitioner developing an agent. We can point out novice problems/traps to avoid, create consciousness that the development of the technology is achievable despite comprehensive and significant challenges, and raise awareness about different ethical issues that may be associated with this technology. We have compiled our experience with deploying conversational systems for daily use in multicultural, multilingual, and intergenerational settings. Additionally, we will give insight on how to scale the proposed solutions.

INTRODUCTION

Before deploying conversational agents as a communication channel between users, a wide range of factors must be considered. Firstly, the task and its level of complexity along with the customer and user's needs, demands, and characteristics. Secondly, the use of available state-of-the-art solutions while appraising for performance, computational costs, and time. Thirdly, checking the amount and quality of the training data and resources for the domain and language. Fourthly, design the best interactions with the chatbot based on the users' environment, and last, but not least, acknowledge and resolve ethical problems that may emerge.

These factors will help determine, among others, if the chatbot can be implemented by using basic rule-based models or more complex systems based on deep learning approaches such as end-to-end systems. Additionally, they will facilitate decisions regarding best interaction with the user, e.g., by voice, text, buttons, or displaying images/videos, and what kind of information it is possible to get from them directly or indirectly. For example, emotions reflected in the voice, face or gestures, noise in the environment, location, number of different users interacting in the same conversation, or user profile: age, gender, preferences, personality, etc. Ideally, the "perfect" agent will interact in a similar fashion to how another human would. In practice however, the effort and cost of creating such a system is substantially high and there may be technology and resource limitations (e.g., different sources of information, lack of training data, system and network latencies, channel limitations), as well as ethical issues to consider (e.g., privacy of data, careful interaction with vulnerable users, addition of cameras to intelligent speakers, use of unbiased algorithms). Although these factors may seem straightforward, they are often undefined and a great deal of time is spent developing solutions that do not fit the task conditions, the users' needs, nor the expectations of the company. Therefore, a careful understanding of the technology, requirements, and thoughtful design are important factors to avoid problems, delays, and frustrations.

Our considerations for some of the most important factors mentioned above have been divided into three broad categories: (a) general considerations, (b) scalability, and (c) ethical aspects. Each section will provide a short introduction to each factor, available solutions, and offer other practical advice. Ultimately, the goal is to teach the practitioner to consistently use these considerations when designing conversational agents, moving beyond the current hype and tendencies for creating real-world conversational systems.

General considerations will cover the aspects to consider for selecting between a rule-based or machine learning based approach, deployment of a speech or graphical interface, and discuss the computational resources and technology required to run the system based on the selected algorithm for the chatbot. Then, this section will also include recommendations on chatbot names and how to handle users who want to meddle with the deployed system.

In scalability, most of the suggestions will relate to the growth of the chatbot from simple and specific uses, towards a more complex system by introducing new modules that take advantage of acquired data along the process as well as feedback from the users.

Lastly, ethical aspects that may arise need to be considered. Aspects associated with data collection, design of notifications, legal matters, and content used for training and providing information especially when the chatbot is intended for minors will be discussed.

GENERAL CONSIDERATIONS

Most commercial chatbots are created as a communication channel between companies and users to resolve a particular problem or task. For instance, providing information about the company or product, collecting data from the user to redirect them to the right human agent, or solving repetitive but simple tasks such as scheduling, buying tickets, or playing music or video. These task-dependent chatbots are usually implemented by means of a reasonable number of rules (i.e., rule-based or decision-tree systems) or information retrieval approaches, the usage of regular expressions to extract relevant information from the answers provided by users, and some predefined and specific dialog flow structure (Cahn 2017; Jurafsky and Martin 2020; López-Cózar et al. 2014; Young 2002).

On the other hand, when the task requires a more unconstrained interaction (Arsovski, Wong, and Cheok 2018), such as answering open questions, chit-chat, handling complaints, or where many possible situations, domains or topics must be handled, then good solutions are currently available through deep learning algorithms or hybrid approaches (combining rules, information retrieval, and deep learning models).

Algorithm selection: rule-based or AI-based solutions

Currently, most chatbots that have won important competitions like the Loebner's prize (AISB 2020), e.g., Mitsuku (Worswick 2019)[1] or Cleverbot[2], consist of thousands of rules to fit the capabilities and knowledge required for users. The main reasons for selecting a ruled- based approach are: (a) limited annotated training data to create a more complex solution (a typical situation when creating a chatbot from scratch for a new domain or task), (b) they are more predictable with respect to the prompts presented to the user and actions to carry out after each interaction, (c) it is an initial solution for data collection and for detecting important user needs and expressions, (d) it can safely attract users' attention to the new communication channel, (e) it can solve very specific tasks, and (f) it can be easily used to redirect the interaction to a human agent after gathering information and some keywords.

On the other hand, recent advances in Natural Language Processing, machine-learning techniques, and large amounts of data have allowed the creation of chatbots based on artificial intelligence (AI) and advanced

Information Retrieval (IR) techniques (Chen et al. 2017). When the data is annotated (Hirschberg and Manning 2015), it can be used for tasks like domain and subtask classification, tagging, detection of entities (e.g., names, quantities, locations and user intents), detection of the important parts of the sentence (i.e., performing natural language understanding, NLU), perform robust training by means of adversarial training techniques (Samanta and Mehta 2017), or to generate and rank diverse and natural answers. Unannotated data can be used in unsupervised algorithms for learning low-dimensional representations of words (e.g., word or sentence embedding vectors) to handle synonyms, paraphrasing, misspellings, infrequent words, sentence similarity, answer ranking, and dialogue context representation (Henderson et al. 2019b). In both cases, the quality of the results highly depends on the amount of data, quality of the annotation, complexity of the task, and careful selection of the algorithm and its configuration. Most frequently, the best approach is to apply transfer learning techniques (Ruder et al. 2019) to optimize the hyper-parameters of large pretrained models on huge amounts of out-of-domain data adapting them using specific domain data (Henderson et al. 2019a).

Finally, although deep learning techniques have led to important improvements in the quality of the chatbots (Nuruzzaman and Hussain 2018), complementary approaches such as rule-based or hybrid algorithms can be used to increase the quality and controllability of the interaction in certain parts of a chatbot like response generation and intent detection (Kurachi, Narukawa, and Hara 2018). It is important to note that while applying machine learning methodologies is desirable, there may be situations where the technology continues to require improvements or may not be fully applicable; therefore it will be better to use not so modern but safer techniques.

User interface design

Another important aspect to consider is the interface used by the chatbot. This is quite important since the interface is directly responsible for providing a first impression to the user, allowing different types of information interchange, and extending or limiting the possibilities for correcting errors. When considering conversational agents there are two major trends: speech-based or graphical-based interfaces.

Speech-based interfaces

Speech is one of the most natural means of communication between humans thus this interface is very popular. With it, designers need to consider the following factors:

Demographic characteristics of the final users: If the chatbot will be used by people of different nationalities (e.g., the employees of a multinational company or users from all over the world), the module for performing the audio transcriptions must deal with different languages, accents, colloquial expressions, mispronunciations, code-switching (i.e., changing languages while speaking) and grammatical errors (Schultz and Kirchhoff 2006, chapter 4, pages 79–90). Most typical solutions for these problems, especially for under-resourced languages or topic-dependent applications, require the use of extended pronunciation and word dictionaries, acoustic and language model adaptations, and dynamic transcription based on geo-localization information (Besacier et al. 2014). This potentially means collecting specific training data for each type of user, region, and task, which will increase the development time and cost.

Cleanliness of the users' environment: In this case, it is necessary to pay attention to the presence, type, and level of noise. For instance, in a factory setting there may be background noise from machinery, for the ambulatory user there may be noise from walking in a crowded place, and for the commuter there may be noise coming from the subway, car, or airplane. In all these cases, noise will be harmful for the speech recognition system (Barker et al. 2017). A few strategies to consider include: (a) cleaning the speech signal by passing it through different types of digital filters (Gupta and Gupta 2016), (b) use a model that suppress the noise and enhance the speech signal (Donahue, Li, and Prabhavalkar 2018; Karjol, Kumar, and Ghosh 2018), (c) perform robust training using speech samples containing a large variety of noises (Ko et al. 2017) or (d) use data augmentation techniques (Park et al. 2019).Unfortunately, the first two strategies could fail if the system needs to recognize speech data in conditions for which the filter or model has not been fine-tuned or there are important mismatches with the type of noise or channel for which the filter or model has been trained. On the other hand, the final two strategies increase the training time required and could burden and slow down the transcription system since the model will require more parameters to learn all the different types of noise. Fortunately, most state-of-the-art speech recognition systems can achieve high transcription rates even in difficult acoustic conditions, they may be adapted to new noisy environments, and handle optimized pipelines to perform fast speech transcription (Chiu et al. 2018).

Switching between topics: Speech recognition systems also need to consider the possibility of switching the language models used to provide the final transcription as topics change during the interaction (Li et al. 2018). In most cases, current cloud-based or state-of-the-art speech recognizers are trained on large amounts of data from

different domains and allow large vocabulary recognition that can be used in most contexts and final users. Besides, these speech recognizers allow designers to specify particular domain terms for which the system apply a higher probability of being transcribed, therefore boosting the accuracy for those terms (Bacchiani et al. 2017). However, if the dynamic switching is complex or it introduces undesirable delays, then the design can be focused on recognizing the most probable or critical terms (keywords) to recognize instead of aiming for a perfect transcription (D'Haro and Banchs 2016; Mani et al. 2020), performing quick adaptations over class-based language models (Vasserman, Haynor, and Aleksic 2016), or running multiple topic-specific recognizers in parallel selecting one at a time based on a topic classifier (although this will be a costly solution).

Handling multiple users: The presence of more than one user speaking at the same time, creates the effect of crosstalk, making it difficult for the system to correctly transcribe the speech. To minimize this problem, end-to-end systems are trained to perform source separation and speech recognition (Seki et al. 2018), intelligent speakers use microphone arrays that can detect changes in the speech signal produced by simultaneous speakers allowing the detection of the direction to the closest speaker and then enhancing the microphone signal for that speaker (Khoubrouy and Hansen 2016). Diarization or speaker recognition algorithms (Shafey, Soltau, and Shafran 2019) could also be used to improve the accuracy and performance of the system. Moreover, most current cloud-based and state-of-the-art speech recognition systems are trained to consider all these issues and offer robust transcriptions trained with data from people with different ages, genders, countries/regions, speaking in different environments, or talking about different topics or using specific jargon (Serizel and Giuliani 2017; Bacchiani et al. 2017).

Visual-based interfaces

When time, user environment, and the quality of the speech transcriptions become critical factors, or when the incorporation of speech recognizers introduces higher costs, high network or running latencies, or depends solely on cloud-based solutions, designers should consider relying primarily on the use of visual interfaces rather than voice-based interfaces (McTear, Callejas, and Griol 2016). In general, visual interfaces have a lot of advantages and allow a high level of flexibility. Although speech input could be faster and more accurate than typing on a mobile device (Ruan et al. 2018), the advantage of using input texts rather than speech transcriptions is that the former can be more easily collected and processed. Below we describe some of the advantages and design considerations.

The simplest visual interface is text message exchange. Its data is easy to collect, and it is not affected by background noise, accent, or signal stability like for speech interfaces. Multiple users within the session is uncommon or can be avoided/detected (i.e., automatic closing of sessions on shared computers, automatic detection of users by keystroke patterns). With a visual interface the detection of a topic or language switching can be performed by applying sliding window analysis (He, Li, and Wu 2017) and grammatical errors can be easily detected (e.g., some apps and browsers already include syntax correction or auto-correction features). Additionally, keyword detection on input texts can be easily implemented using basic regular expressions or more advanced approaches like conditional random fields or deep neural taggers (Andor et al. 2016; Huang, Xu, and Yu 2015; Jurafsky and Martin 2020; Sarkar 2018).

Visual interfaces permit the use of additional elements such as buttons, option lists, interactive maps, tables, highlights, or graphics. These elements can be helpful in reducing chance for error and for collection of critical or private information. Video or images may be added to facilitate product description (e.g., items to be bought online) which can provide a faster, or at least complementary, interaction than using voice or text alone. Finally, since the amount of textual data that can be used from existing resources/repositories is larger than for speech data, it is possible to train more complex systems for dealing with the different tasks required for text processing.

On the other hand, designers still must overcome a few problematic obstacles which continue to be the basis of ongoing research. These include misspelling, acronyms, emoticons and jargon usage, handling wrong auto-completes, use of short sentences which imply a higher number of steps to get to the information needed to perform the task (Gupta and Joshi 2017; Sproat and Jaitly 2016). Besides, text-based interfaces are not always practical or even or even allowed in some settings; driving, handling machinery, or while interacting with other users. In these cases, a well-designed speech or simplistic graphical-based application is more convenient and safer.

Finally, designers also need to consider the age and physical capabilities of their users. Children may not be able to read or write, elderly people may have difficulty reading or typing on a device, and people may have other difficulties to be considered (Følstad and Brandtzæg 2017; Lister et al. 2020; Yuan et al. 2019). In such cases, it may be more suitable to design a chatbot with a mixture of graphical elements and speech messages.

Computational resources

Another design element to consider is real-time response. Overall system latency must be reduced to make the interaction more comfortable and natural; thus, the selected technical solutions must be efficient in terms of memory, processing time and physical resources. To address this problem, it is important to select proper algorithms and architecture. Below we provide some of the most important issues to consider and their common solutions.

Computational load due to selected algorithms: Currently, Deep Neural Networks (DNNs) are the main trend due to their excellent performance but they tend to demand high technical and computational resources which can be a limitation for mobile or embedded systems. In such as case, using cloud services is an alternative bearing in mind that a network connection is needed, data usage needs to be reduced, and there could be privacy issues. Alternatively, in case connection, speed, or model size are important factors, information retrieval or rule-based approaches can be considered as good options for some tasks like classification, understanding or generation. Finally, machine learning designers can help by reducing the size of the models (e.g., reducing the number of layers, applying distillation techniques or reducing the vocabulary size), reducing the number of stored records or rules in databases, and reducing the number of topics to classify or handle (Cheng et al. 2017; Polino, Pascanu, and Alistarh 2018).

Delays due to network and dialog flow processes: In this case, the conversational system and network latency is high due to performing third-party API calls, database access, or when high demanding resources are needed (e.g., processing long videos or long text utterances). The first design element to consider is to provide feedback to the waiting user. A basic mechanism for feedback is to use dancing dots or an hourglass rotating icon in a graphical interface, or in other forms like music or a speech-based information message given at regular intervals (Cohen et al. 2004; Li and Chen 2019). Complementarily, it is possible to reduce system latencies by detecting the most common situations or user requests from logs of previous conversations, which then makes it possible to prepare responses in advance and create a flow that could suit most of the users and their needs (Ondáš et al. 2018).

A more complex solution would be to modify the dialog flow and process the actions that take longer first, or closer to the beginning of the dialogue, while concurrently performing other faster actions (i.e., retrieve a higher number of results from a server by using incomplete or a priori data; using geo-location or typical queries, and then perform a local filtering of those results, getting a summary of the latest news while supplying information about the expected weather).

Finally, some promising research focuses on incrementally training dialogue systems using reinforcement learning techniques where the dialogue management module learns to predict user actions by anticipating their responses (Khouzaimi, Laroche, and Lefèvre 2018) while being rewarded if the task is completed in the least number of steps. This way, the final conversational system will dynamically adapt its dialog flow in such a way that the interaction with the user is the shortest one to complete the task.

Selection of technology providers

Designers also need to consider the criteria for selecting the platform, technology, and tools provided for deploying, generating, and evaluating the proposed agent.

Selection of the design platform: Most current service providers or chatbot creation platforms provide very intuitive graphical interfaces to design the dialog flow, predefined prompts and responses, and models to perform language understanding and grammar parsing. They also allow for the combination of text and graphical interface elements, and can be adapted for reuse across multiple languages, and communication channels (e.g., Twitter, Telegram, WhatsApp, Facebook messenger or Slack), and offer powerful services for performing understanding and intent classification (Liu et al. 2019a).

However, designers should consider whether the platform provides predefined procedures for handling typical errors or situations. For example, repeating a previous prompt, retreating in the dialog flow (in case the user clicks on the wrong button or misunderstands the question), remaining in the same inquiry in case the speech transcription contains errors or its confidence value is below a given threshold, detecting implicit changes of topic, or handling the situation when the chatbot provides a wrong answer to the previous turn or dialog context. In these cases, a good platform will provide pre-set procedures and prompts for seamlessly handling these errors creating this way a better user experience, while not increasing the designers' effort and time in development. Additional considerations and desirable properties can be found in (Di Prospero et al. 2017; Kostelník et al. 2019; Ray and Mathew 2018).

Selection of the service provider: Although the selection of a big tech provider could be tempting, a smaller or local provider may provide a better service based on a lower latency, technical support in the local language, and adapted tools/models for the idiosyncrasies or characteristics of the end users. Designers need to consider how the provider handles the data, the security of their system

(e.g., using containers, access permissions), load balancing and additional resources during peak hours, as well as management tools to control and visualize usage of the system, resources, performance, and handling of problems.

Other practical recommendations

In addition to all the aforementioned considerations, there are a few more recommendations that will be particularly useful for those people new to this type of technology design.

System prompts presented to the final users: When creating a solo voice-based application, it is very important to use short, precise, and very well-explained prompts, because the lack of a screen makes it very easy for the user to get lost. If this happens, both the conversational agent and the user will enter into a loop where they cannot understand each other or find a clear exit, therefore the user will not use the system again. Special consideration should be given to the prompts used to indicate the system cannot understand the user. In this case, it may be necessary to ask the user for a rephrase, provide a successful contextualize example, or provide an alternative strategy like check for spelling.

Interaction with non-cooperative users: Similarly, it is important to foresee the interaction with trolls, i.e., people who ask strange things or try to drive the chatbot into complex or troublesome situations. Thus, it is useful to prepare beforehand a battery of phrases to ask the user to change/limit this behavior.

Selection of chatbot's name: Another aspect is to correctly choose the name of the chatbot (or skill, as in the case of Alexa) looking especially for a short name that is easy to pronounce and if possible, provide insight on the scope or topic. Consider that 75% of the time, voice search responses will come from the first three results (Enge 2019). This will ease the process of using and calling it, while consequently increasing market position or value. For an interesting article with additional factors and recommendations for creating good chatbots based on the experience of more than 200 users, please consider reading (Følstad and Brandtzaeg 2020).

SCALABILITY CONSIDERATIONS

According to Gartner's predictions (Kaczorowska-Spychalska 2019), almost 85% of customers' relationship with enterprises will be managed interactively through a conversational assistant by 2020. This statistic shows the relevance for this trend and the importance of creating a good AI-based assistant that could make the difference in such a competitive market. To ensure chatbots carry out their task appropriately, it is essential to know the level of complexity a chatbot can handle, i.e., what is the ultimate goal for the chatbot, and its scalability. Below we provide some general ideas for designing and extending a chatbot.

Keep it simple: The first thing a designer need to consider is that a unimodal chatbot (e.g., text-based) focused on a reduced/specific number of tasks will be easier to handle and scale over time than a multimodal one (e.g., voice, text and images) that deals with varied/numerous types of tasks for different departments or sections within a corporation. Although this seems straightforward, the reality is that there is a greater trend to create complex agents prone to fail rather than elaborating simple but efficient and scalable ones.

One of our successful chatbots, as an example, was developed to assist new employees by providing them with answers about typical administrative procedures or information, or to answer basic questions like "where can I eat at the office?". The proposed solution was to use a closed-domain decision tree to drive the conversation over a specific path depending on the customers' answers to predefined questions posed by the chatbot. The advantage is that it allowed full control over the conversation because customers could choose from fixed answers, via a set of buttons (with the predefined answers), which in turn were translated into specific values in the decision tree. This simple implementation allowed the chatbot to be quickly adjusted to different situations, which would otherwise not be possible in an open-domain task.

The previous example highlights that it is not necessary to create a complex conversational agent to accomplish some tasks effectively. In fact, starting with simple models and then updating/escalating their functionalities, by adding new decision trees to augment the domain understanding and number of tasks, allows a faster deployment and evaluation than creating a more complex system from the start. Later in our development, we had the opportunity of integrating external components to quickly improve the chatbot, by increasing its personalization and ultimately customer satisfaction, while avoiding developing those features from scratch.

Use hybrid systems: As commented in Section 2.1 (combining rule-based or AI-based algorithms), the use of hybrid systems allow the creation of more complex systems without requiring too much data for training or moving into complex end-to-end systems (Song et al. 2018), and are therefore useful when discussing scalability. In this context, the use of statistical approaches can be left for specific components or modules where enough data for training a model is available, while using simple fixed solutions when there is not enough data or a more controlled behavior is desirable. For instance, it is possible to create a closed-domain decision tree with fixed answers for

well-known questions, while a machine learning module could be used to detect closer user paraphrases to predefined questions, or to detect the users' intents in those questions even if the input question contains misspellings or errors in the transcription (Tammewar et al. 2018). The advantage for this solution is that although the questions can change their grammatical structure and words, the semantic representations of the sentences will be close enough to the known ones, permitting the system to answer the new questions with the same predefined question/answer set (i.e., helping to scale the system and make it more robust).

Introducing variability for system responses: If designers want to include a certain level of variability to a set of fixed or limited responses, allowing the adaptation of the chatbot to different kind of users, situation, or providing a more personalized experience, they can consider the incorporation of an NLG (Natural Language Generation) module which can be used together with different templates (i.e., fixed sentences or words with some given empty slots that will be completed according to the dialogue state, e.g., "you have bought <NUMBER_TICKETS> tickets. Thanks for your purchase") for each response allowing the selection of one of those templates at runtime.

A couple of important items to note regarding the proposed strategy for variability include: (a) the number of templates may not scale in the long term (i.e., there could be too many templates or rules that is difficult to update or maintain them), and (b) post-processing of the generated sentences must be done to avoid grammatical errors that will produce a bad impression in the final users, e.g., in the example above, if the number of tickets is just one, then the following word after the slot should not be tickets in plural but ticket in singular.

Alternatively, it is possible to rely on deep learning approaches such as generator-based systems in order to provide additional freedom in the sentences (Jurafsky and Martin 2020 – Chapter 26; Ramesh et al. 2017). Recent transformer-based models (Vaswani et al. 2017) using deep neural models such as BERT (Devlin et al. 2018), GPT-2 (Radford et al. 2019) or the more recent GPT-3 (Brown et al. 2020) have shown excellent results on free-text generation and transfer learning. These models automatically generate high quality and syntactically correct sentences across multiple domains and potentially allowing scalability to dialogue systems. However, their internal algorithms are still in their initial stage of research therefore caution is advised. In this case, more conditional generation algorithms, such as PPLM (Dathathri et al. 2019) or CTRL (Keskar et al. 2019), may be required to allow for knowledge-based, engaging and domain-relevant generated sentences.

ETHICAL CONSIDERATIONS

Ethical consideration within the field of artificial intelligence is increasingly relevant. The technology has advanced at such a rapid pace, that the public has had difficulty keeping up with its impact, and not surprisingly in dealing with the unethical and inappropriate use of it. Sadly, some AI developers are taking advantage of the lack of regulation to perpetrate their negative ambitions. Below, several issues and potential solutions within the context of conversational agents are presented. The main goal is to underscore ethics, transparency, and responsible use of the technology.

The problem behind data scarcity: It is well known that the lack of data directly hurdles the performance and improvement of conversational systems. Firstly, the lack of conversational data impedes systems from achieving the level of precision recently accomplished in areas such as voice recognition, image classification or object detection, where thousands of hours of speech or millions of annotated images are available, and are often used in competitions allowing swift improvements that were not possible a few years ago. In dialogue systems, such datasets are scarce, their size limited (i.e., some are just a few thousand interactions), usually collected in non-natural environments (i.e., controlled topics and content), and often contain data that is not relevant to more general domains (Serban et al. 2015). In addition, it is often difficult to automatically assess the quality, level of interaction, engagement or naturalness of conversations between humans and other humans or chatbots that could be released with the goal of training more robust and scalable systems. Recent metrics have been proposed (D'Haro et al. 2019; Tao et al. 2018) but this is still an ongoing topic of research that will bring important applicability in the near future.

While some interesting solutions are available e.g., by using data augmentation techniques (Du and Black 2018; Hou et al. 2018), performing back-translations to different languages (Fadaee, Bisazza, and Monz 2017), or using automatic paraphrasing systems (Wieting and Gimpel 2017)[3], the reality is that such techniques are not entirely able to recreate the high level of variability observed in human interaction. Moreover, since the process of creating real datasets could be slow and expensive, it is important to create mechanisms that can automatically anonymize collected data allowing data sharing among the research community.

Data anonymization issues: Although many anonymization algorithms exist (Li and Qin 2017) based on named-entity recognition, coreference resolution, collocations, or keyword spotting. It is important to consider that these algorithms may not be enough to remove all personal data due to the dynamic nature of the human

dialogue (e.g., high variation of named entities over time, introduction or use of new words, jargon or expressions among people or domains). Besides, anonymization algorithms could still allow the collection of personal information in an indirect way. For instance, demographic information could be inferred from interaction logs of user activities (Henderson et al. 2018). Therefore, data anonymity is relevant when considering solutions for data scarcity. Without it, data would be even more limited, making the problem of scarcity worse and a slower growth for general conversational system performance.

Issues with data collection and quality: As mentioned above, when designing and deploying a chatbot, it is common practice to gather data from a similar domain available online or from previously, and probably anonymized, data from customer service logs. In time, as the first prototypes of the chatbots are tested and deployed, new data emerges allowing the developer to improve the functionality of the chatbot. This assortment of data may give rise to ethical dilemmas regarding private data. For example, several tech companies (Vanian and Pressman 2019) record all interactions with users in order to improve their algorithms, personalize, annotate new interactions, or detect new topics or interests, but, in some cases, this data collection is carried out when users are not using the application (e.g., continuously listening to the environment for improving noise cancelation or wake-up algorithms). Designers need to consider that the use of private data may have ethical implications and the question of whether users should be informed about how this data is collected, handled, stored, or used becomes important. Particularly, when using collected data from external sources, it is important to assess its quality in terms of its closeness to the domain data, vocabulary usage, avoiding gender, race, or other kind of biasing, and avoiding usage of improper language (Henderson et al. 2018).

Control over collected data and users: One important goal for private data management should be to retain customers' trust by allowing them control over what data is accessed, shared or deleted (i.e., right to be forgotten). Fortunately, there are a few ways by which the dilemma of private data use can be averted: (a) users may get secure online access to their stored information, unimportant information may be periodically and automatically removed, and adaptation mechanisms may be used where some new interactions are used to keep the system updated and relevant, while the rest is discarded to avoid biasing the system towards atypical situations, and (b) some companies reduce ethical problems by asking specific users to opt in as a control group. These users are aware that their interactions are highly recorded and go under detailed contracts to reduce legal and ethical issues but allow their profiles to be used to improve the interactions for similar users.

Issues when doing dynamic adaptation to users: When considering conversational agents that adapt conversations based on the estimated users' mood, estimations may not be explicitly extrapolated from the user, but discreetly performed and executed by algorithms that analyze text content, tone or speed in the voice, face or posture analysis. While the estimation and dialogue adaptation without informing the user is a common procedure to keep the fluency and naturalness of the interaction and reduce the sense of invasiveness, designers need to keep in mind that the estimation could be erroneous (e.g., due to biasing or class imbalance in the training data for the emotion classification) and therefore be careful in its use to avoid the system updating the user's profile using a wrong estimation (Kang et al. 2018; Liu et al. 2019b). In addition, designers can deploy algorithms that could explicitly or implicitly confirm the estimation in a similar way as with speech recognition results with low confidence estimation (Litman, Walker, and Kearns 1999).

System notifications: System notifications are not exempt from creating ethical issues. Personal information may be obtained by third parties when the chatbot provides an audible notification due to a reminder about an item in the personal agenda. Solutions for these kinds of problems go from careful understanding of the available modalities and their capabilities; for instance, scanning the environment in order to decide the best time, modality or mechanism to present the information, or using explicit configuration questions about the user's availability for getting notifications on a given period of time, for urgent matters only, etc.; in any case, it is important to avoid asking unnecessary questions or handling compromising information. Even chatbots need to be designed to comply with the current personal data protection regulations (GDPR) (Saglam and Nurse 2020), and also to inform the user that the interaction is being done with a chatbot instead of with a human (Ischen et al. 2019).

Proper language, data curation, biasing, and awareness: Lastly, designers need to consider the characteristics of the end-users when defining the language used by the chatbot and the content to be share. For instance, when interacting with minors, proper language is important. Therefore a careful selection of the sentences/prompts to be presented to the users is important; however, current state-of-the-art generative deep neural approaches, although very promising as mentioned above, could generate uncontrolled sentences in the case that the training data and vocabulary are not properly selected, unbiased, or that the data that is used to dynamically retrain the models used by the chatbot is not carefully selected. Therefore, it is possible that the chatbot presents sentences that are not appropriate for the final user. Consider the famous case of Tay, the Microsoft chatbot that went uncontrolled after a short time online (Wolf, Miller, and Grodzinsky 2017). Thus,

using retrieval approaches over curated data or controlling generative model outputs with hybrid models are possible solutions. Alternatively, using rule-based models is a safer solution because system messages are kept under control, at least for certain domains, type of users, or tasks. For some additional thoughts and considerations regarding ethics for chatbots please check (Reddy 2017; Shanbhag 2020) and (Ruane, Birhane, and Ventresque 2019).

SOME THOUGHTS FROM OUR EXPERIENCE: DESIGN, ENGAGINGNESS AND PERSONALITY

As discussed above, different factors including proper design, scalability, and ethics should be considered when creating, designing, and running a good conversational agent. In this section, we would like to share some of our recent developments from the perspective of a start-up, born in an academic environment, which navigated us through a very intense but enriching experience. These experiences may be useful and inspiring to students, newcomers, or industrial partners willing to move into using chatbots for their businesses.

The importance of a correct design

One of our first challenges was to merge our academic/technical knowledge and applied business knowledge to reach the general public. We began by analyzing the market demand and realized many businesses wanted to have these kinds of agents to attract users and stay relevant in the marketplace. The problem was many did not have any kind of data to start with, nor an idea of what kind of user needs they wanted to meet or fulfill. Therefore, the need and task had to be defined. Next, it was important to have an understanding of the technologies, requirements, quality and size of the available training data (e.g., phone-call logs, surveys or social media channels) and directly inquire from the base, whether it be the company section or the intended users, regarding their actual requests to avoid delays, misunderstandings, and falling into the tech hype.

To address the data scarcity problem, an initial chatbot was designed to collect information quickly and easily from users while providing basic services like answers to general or common questions. This was made possible by using online development platforms with pretrained models or rule-based systems. The use of graphical buttons for collecting answers made it easy to check later what kind of answers were more common, allowing the transition to more complex solutions using natural language sentences. Once the first prototype was complete, we created a control group with users that were motivated/enthusiastic to evaluate the system and that provided us with objective feedback to improve the chatbot.

From real client experience, we concluded that most people were willing to use the chatbot when the conversation was fully guided (i.e., system-initiative) with simple instructions and prompts. After the initial experience, clients readily transitioned towards a chatbot that could give them more freedom to guide the dialogue. This is typical behavior because novice users do not know the system and its capabilities and therefore act more apprehensively; giving them too much information blurs the message to be transmitted and complicates the data collection, especially if the user gets tired of the interaction and finishes it even before delivery of all the required information or the task has been accomplished. Users would rather adapt to continuous improvements than rely on a chatbot that has more capabilities but fails to meet their needs (uncanny valley). A bad first impression is worse than using a limited but useful chatbot.

The importance of the interface

One of our successful developments was an onboarding chatbot for a Japanese multinational. The company wanted to automatically provide useful information to newly hired employees during their first 3 months of work and collect feedback on their initial impressions about the chatbot. Requirements made by the company for this project included making data collection for future analysis automatic (i.e., in an innovative and visual manner), and that the definition of the questions and responses could be done by a layman employee. At the same time, the interaction with the users quick and short.

To address these requests, we focused on creating a web-based application and management interface allowing users to use their mobiles, and system administrators to manage and edit the dialog flow design, visualize the logs of the chatbot, and create new or edit questions and responses.

For the final users, the use of a web application made the interaction with the final users easier and quicker because mobile terminals and messaging applications are widespread; users are accustomed to sending and receiving text, image, audio, documents, GIFs, geolocation, notifications etc. reducing the entry barrier. In addition, users did not have to download and install a new ad-hoc application (something that most users are reluctant due to restrictions in the mobile capabilities, security concerns about new applications, or simple tiredness due to bad past experiences). On the other hand, most current design platforms allow the reuse of previously installed apps such as Telegram, WhatsApp, Facebook Messenger, WeChat, or Slack. Using these pre-installed apps also facilitates the

process of showing messages with images, audios, gifs, sending documents or web links, and exploits push notifications to bring users' attention and curiosity when there is a new message.

Moreover, the use of mobiles allowed us to accelerate the sending process. When our chatbot sent a scheduled message to users, this process was done instantaneously and almost without human intervention. The integration with commercial messaging applications made it possible to save the conversations and carry them out on different devices, e.g., computer, mobile or tablet. Employees were therefore able to read the tips on their office PC, at home using a tablet, or during their commute via their mobile phones. The result was that our chatbot was greatly valued by the employees.

In addition, in order to guide the interactions with the user (i.e., system-initiative), we employed a simple technique: the chatbot sent a message (or several consecutive ones) together with some predefined answers shown using buttons allowing to acquire a unique answer (or several ones) reducing possible errors in typing and allowing a soft and guided interaction with the user. Internally, and for management purposes, we designed a predefined decision tree where the nodes were questions to be presented to the users; based on their answers, the conversation went through a specific and personalized path. With this simple design technique, the users' perception was that interactions were customized and intelligent, requiring minimal effort from them, which meant that our clients assessed the chatbot very positively. Moreover, another advantage of creating the decision tree was that we could use a visual interface to show it, allowing chatbot administrators to customize the decision trees by dragging icons and joining them with lines, specifying the messages that they wanted to ask or tell users, and even uploading multimedia files that they wanted to show.

Another functionality that our client found very important was to automatically send reminder messages. In this case, from the platform, by just clicking on a button in the interface and scheduling the time in minutes/hours/days, the information was saved and sent automatically to all users subscribed to the chatbot. Finally, the administrator could specify when to repeat or resend the message, to send a given message to all users or only to a specific set of people, groups or departments, allowing a certain level of personalization avoiding spamming users.

The importance of personality in chatbots

Finally, we want to share two successful experiences we had with the development of two chatbots intended for two different but sensitive groups and where the personality of the chatbots significantly contribute to their success.

The first chatbot was intended for a young adult (ages 18–26) audience in Spain. In this case, the chatbot was utilized to forward important messages, whether they be related to work or school, but send the messages in a way that would make them want to read them. If the message was regarding a specific task, then the user could also send a notification of the completed task (two-way communication chatbot). The problem to be solved in this project was clear: how to make the target audience interested in reading notifications? People within this age group are sometimes reluctant to read communications by traditional means (e-mail, paper, SMS, etc.), this leads to important messages being neglected. Given their demographics and the problem to be resolved, we considered that we could gain and keep these young people's attention by means of designing the chatbot with an engaging personality.

Personality is a feature that can potentially enhance a product/chatbot (Duijst 2017; Smestad 2018). If it is unique enough, consumers will remember the personality just as they do when they think of a person (bearing in mind the opposite is also true). Therefore, the language and demeanor of the chatbot must be chosen carefully. The strategy used here to attract these young people was particularly risky. We provided the chatbot with a name and surname and gave it personality as if it were a live robot. It could behave at times as a well-mannered, intellectual person or as an ironic, insolent, thug, and could formulate jokes and make bold statements. Keeping the chatbot respectful was of course very important, so we carefully chose words and tone that would be appropriate for the desired personality. Besides, we wanted to strengthen the relationship between the chatbot and the user by using some few strategies: calling the user by name, sending messages giving the perception the chatbot was a great friend, telling users that the robot missed them, sending the user animated gifs or writings that would show affection, and finally adding components of humor and irony to make the conversations more enjoyable.

As a result, we found that these young people read most of the communications and read the messages in their entirety. This was evident from the results by the automatic evaluation surveys collected by the chatbot. In addition, when we analyzed the logs, we found that, in general, users became closer with the chatbot by remembering it and calling it by its name, and waited until the end of the interaction because of the funny ways the chatbot said goodbye to them.

The second chatbot was intended for children between 3 and 12 years old. This time, we were involved in the design and programming of an Alexa type chatbot. The task involved the development of a question and answer voice-based game that involved conversations about the children's favorite cartoons or movies, whether it be about the plot, the characters, basically anything related to them.

Devoting to this type of audience required consideration of numerous ethical aspects and exceptional care in the use of language. Two goals were set: (a) to ease interaction so that any child could play with our game without difficulty, and (b) to try to make the children feel part of the project.

To achieve the first goal, we designed a very guided interaction, with use of colloquial expressions suitable for such ages, by increasing the use of diminutives in certain words and applying a loving tone to most of the prompts. Sentences were also shortened and simplified so that children could understand them. For the second goal, questions and answers of the game were about cartoons and movies they watched regularly. Knowing the topic of the questions made the users more likely to succeed and therefore feel better about themselves and their capabilities.

Building confidence is an important and positive aspect to develop in young children. We took our chatbot one step further and allowed children to participate by sending their questions via email or a web page form, some with the help and permission of their parents, while Alexa thanked them personally with their name before reading the question and providing the corresponding answer. This made the children feel they could make the game their own and be able to modify how they wanted to interact with the system each day. However, to avoid engaging them for too long or distract them from other activities, we limited the playtime to a few hours in the afternoon, and the number of questions to 5 per day. We also considered here that it was better to create a use routine, however short, rather than have users use the chatbot 1 day and then not use it anymore or much later. The result was very positive: in 180 days we had more than 2.1 K downloads in Alexa, and more than 1.1 K independent users. For several weeks, our skill was ranked top in the children's section of the Alexa Skills Store for Spain.

CONCLUSIONS AND PERSPECTIVES

In this paper we have presented a set of aspects and recommendations that conversational agent designers should consider when developing a chatbot. Starting from making a clear definition of the problem that the chatbot will address, the kind of users, environments, available resources for training and testing, and careful considerations of the technologies and their advantages and limitations. We looked at scalability and ethical considerations and went through some successful and practical examples where careful considerations of previously mentioned design issues were applied.

An ongoing, substantial challenge that conversational agents are facing is making conversations more natural and engaging. To help the matter, enterprises focus on giving persona and personality characteristics to conversational agents (Dinan et al. 2019; Katz 2019). We successfully experienced the advantage of implementing these characteristics by allowing our bot to answer with a more informal style than what conversational assistants usually do. We carefully designed the interaction but knew in advance the context in which our chatbot was going to work. When this is not possible, it is important for a conversational agent to be able to correctly personalize the answers and not simply use sentences like "I don't understand" or "I don't know".

To incorporate attractive personalities and allow better engagements, we expect most platform providers to incorporate default templates and quick mechanisms to create more personalized and attractive chatbots. For example, we expect that given a question like "How much does the Everest measure" would make the chatbot guess the intent of the user question and provide an exact measure of its height 8,848 meters, or to provide a more typical human response like "around 8000 meters." In this way, the chatbot will answer not like an encyclopedia but like a person would. If a person were to ask a question regarding musical preference and mentioned he or she likes the Beatles, the chatbot could instead of starting a Wikipedia-like summary, mention some particular interest about some of their songs as part of its encoded personality. For these kinds of situations, solutions start with post-processing modules for the NLG component in order to modify how exact must be the answers based on multimodal inputs such as user's profile, environment, chatbot personality, and content of the knowledge base (See et al. 2019). Later, deep learning and reinforcement learning methods can be applied to adapt more efficiently the chatbot to eventual environment alterations that may occur during the interactions.

Another important challenge is the development of automatic evaluation protocols and metrics that can be used to reduce the current approach of asking humans to provide their subjective perception and detect mistakes (Deriu et al. 2019; Kong-Vega et al. 2019). The limitation of having objective metrics directly focused for dialogue systems slow down the use of current state-of-the-art artificial intelligence algorithms. Although interesting metrics and models are being proposed with high correlation to human users (Deriu et al. 2019; D'Haro et al. 2019; Mehri and Eskenazi 2020; Tao et al. 2018; Yuwono, Wu, and D'Haro 2019; Zhang et al. 2020), this is still an area of open research.

An interesting topic of research and future development is the active perception of surrounding context to provide a more natural experience with the conversational agent (Kocielnik et al. 2018). In this case, the information to the chatbot will not depend only on the data obtained from the user or from earlier conversations, but also on the environment and user's lifestyle. In order to move towards

this direction, IoT technology can be used to collect data from the environment through different kinds of sensors, allowing the chatbot to recognize and act according to the changes that happen around the user; for example, agents will know what the room temperature is, which can be used to make the agent more proactive to start a task-oriented sub-dialogue like this: "it is cold here, do you want me to turn the thermostat up? Or in a chit-chat context to express feelings like: "this made me feel uncomfortable, like the temperature in this room. It's too cold."

In time, the aim is to implement AI systems in more widespread environments, such as health, education, research, and even more challenging ones like legal environments. These assistants can be a great tool to search hundreds of thousands of documents for certain specific data in a few seconds, to summarize and highlight important parts of a document, or to automatically create a battery of questions about an academic book in order to help users to study more efficiently. In addition, this type of assistant will be able to adapt to special needs users such as blind or deaf people. Therefore, clear understanding, adaptation and scalability are important aspects to consider and research on. We hope that this document has provided some insights into the challenges, design requirements and experiences that could help to move chatbots towards their next generation.

ACKNOWLEDGMENTS

We specially thank the reviewers for their important and insightful comments to improve this paper. In addition, we want to thank Erikka Baehring and Kheng Hui Yeo for their deep proof-reading and contributions to make this paper clear and suitable for a wider audience. This paper has been supported by the following projects: AMIC (MINECO, TIN2017-85854-C4-4-R) and CAVIAR (MINECO, TEC2017- 84593-C2-1-R) partially funded by the European Union.

ENDNOTES

[1] https://www.pandorabots.com/mitsuku/

[2] https://www.cleverbot.com/

[3] https://github.com/vsuthichai/paraphraser

REFERENCES

AISB. 2020. The Society for the Study of Artificial Intelligence and Simulation of Behaviour. Official website for the Loebner Prize. https://aisb.org.uk

Andor, D., C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins. 2016. Globally Normalized Transition-based Neural Networks. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2442–52.

Arsovski, S., S. Wong, and A. D. Cheok. 2018. "Open-Domain Neural Conversational Agents: The Step Towards Artificial General Intelligence." *International Journal of Advanced Computer Science and Applications* 9(6): 402–8.

Bacchiani, M., F. Beaufays, A. Gruenstein, P. Moreno, J. Schalkwyk, T. Strohman, and H. Zen. 2017. "Speech Research at Google to Enable Universal Speech Interfaces." In *New Era for Robust Speech Recognition*, 385–99. Cham: Springer.

Barker, J. P., R. Marxer, E. Vincent, and S. Watanabe. 2017. "The CHiME Challenges: Robust Speech Recognition in Everyday Environments." In *New Era for Robust Speech Recognition*, 327–44. Cham: Springer.

Besacier, L., E. Barnard, A. Karpov, and T. Schultz. 2014. "Automatic Speech Recognition for Under-resourced Languages: A Survey." *Speech Communication* 56: 85–100.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal. 2020. Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

Cahn, J. 2017. University of Pennsylvania School of Engineering and Applied Science. Department of Computer and Information Science. Senior Thesis. CHATBOT: Architecture, Design, and Development.

Chen, H., X. Liu, D. Yin, and J. Tang. 2017. "A Survey on Dialogue Systems: Recent Advances and New Frontiers." *Acm Sigkdd Explorations Newsletter* 19(2): 25–35.

Cheng, Y., D. Wang, P. Zhou, and T. Zhang. 2017. A Survey of Model Compression and Acceleration for Deep Neural Networks. arXiv preprint.arXiv:1710.09282.

Chiu, C. C., T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, and N. Jaitly. 2018. "State-of-the-Art Speech Recognition with Sequence-to-Sequence Models." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4774–8). IEEE.

Cohen, M. H., M. H. Cohen, J. P. Giangola, and J. Balogh. 2004. *Voice User Interface Design*. Addison- Wesley Professional.

D'Haro, L. F., R. E. Banchs, C. Hori, and H. Li. 2019. "Automatic Evaluation of End-To-End Dialog Systems with Adequacy-Fluency Metrics." *Computer Speech and Language* 55: 200–15.

D'Haro, L. F., and R. E. Banchs. 2016. "Automatic Correction of ASR Outputs by Using Machine Translation." *Proceedings Interspeech* 2016: 3469–73.

Dathathri, S., A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. International Conference on Learning Representations (ICLR).

Deriu, J., A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak. 2019. Survey on Evaluation Methods for Dialogue Systems. arXiv preprint arXiv:1905.04071.

Devlin, J., M.W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Dinan, E., V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, and S. Prabhumoye. 2019. The Second Conversational Intelligence Challenge (convai2). arXiv preprint arXiv:1902.00098.

Di Prospero, A., N. Norouzi, M. Fokaefs, and M. Litoiu. 2017. "Chatbots as Assistants: An Architectural Framework." In Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering (pp. 76–86).

Donahue, C., B. Li, and R. Prabhavalkar. 2018. "Exploring Speech Enhancement with Generative Adversarial Networks for Robust

Speech Recognition." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5024–8). IEEE.

Du, W., and A. W. Black. 2018. Data Augmentation for Neural Online Chat Response Selection. Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI, pp. 52–8.

Duijst, D. 2017. Can We Improve the User Experience of Chatbots with Personalisation? Master's Thesis. University of Amsterdam.

Enge, E. 2019. November 19. Your Roadmap to Featured Snippets. Perficient Digital (Blog entry). Available at https://www.perficient.com/insights/research-hub/featured-snippets-guide

Fadaee, M., A. Bisazza, and C. Monz. 2017. Data Augmentation for Low-Resource Neural Machine Translation. arXiv preprint arXiv:1705.00440.

Følstad, A., and P.B. Brandtzaeg. 2020. "Users' experiences with chatbots: Findings from a questionnaire study." *Quality and User Experience* 5: 1–14.

Følstad, A., and P.B. Brandtzæg. 2017. "Chatbots and the New World of HCI." *Interactions* 24(4): 38–42.

Gupta, I., and N. Joshi. 2017. "Tweet Normalization: A Knowledge-based Approach." In 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS) (pp. 157–62). IEEE.

Gupta, K., and D. Gupta. 2016."An Analysis on LPC, RASTA and MFCC Techniques in Automatic Speech Recognition System." In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), (pp. 493–7). IEEE.

He, J., L. Li, and X. Wu. 2017. "A Self-Adaptive Sliding Window-Based Topic Model for Non- Uniform Texts." In 2017 IEEE International Conference on Data Mining (ICDM) (pp. 147–56). IEEE.

Henderson, M., I. Vulić, D. Gerz, I. Casanueva, P. Budzianowski, S. Coope, G. Spithourakis, T. H. Wen, N. Mrkšić, and P.H. Su. 2019a. Training Neural Response Selection for Task-Oriented Dialogue Systems. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pages 5392–404. Florence, Italy, July 28 - August 2, 2019.

Henderson, M., I. Casanueva, N. Mrkšić, P. H. Su, and I. Vulić. 2019b. ConveRT: Efficient and Accurate Conversational Representations From Transformers. Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2161–74.

Henderson, P., K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau. 2018. "Ethical Challenges in Data-Driven Dialogue Systems." In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 123–9.

Hirschberg, J., and C. D. Manning. 2015. "Advances in Natural Language Processing." *Science* 349(6245): 261–6.

Hou, Y., Y. Liu, W. Che, and T. Liu. 2018. Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding. Proceedings of the 27th International Conference on Computational Linguistics (COLING), pp. 1234–45.

Huang, Z., W. Xu, and K. Yu 2015. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Ischen, C., T. Araujo, H. Voorveld, G. van Noort, and E. Smit. 2019. "Privacy Concerns in Chatbot Interactions." In *International Workshop on Chatbot Research and Design*, 34–48. Cham: Springer.

Jurafsky, D., and J. H. Martin. 2020. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*. 3rd Edition. Available online at https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Kaczorowska-Spychalska, D. 2019. "How Chatbots Influence Marketing." *Management* 23(1): 251–70. https://doi.org/10.2478/manment-2019-0015

Kang, Y., Y. Zhang, J. K. Kummerfeld, L. Tang, and J. Mars. 2018. "Data Collection for Dialogue System: A Startup Perspective." In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers) (pp. 33–40).

Karjol, P., M. A. Kumar, and P.K. Ghosh. 2018. "Speech Enhancement Using Multiple Deep Neural Networks." In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5049–52). IEEE.

Katz, S. 2019. The Ultimate Guide to Chatbot Personality. Chatbots Magazine. (Blog entry). Available online at https://chatbotsmagazine.com/the-ultimate-guide-to-chatbot-personality-b9665ab5e99d

Keskar, N. S., B. McCann, L. R. Varshney, C. Xiong, and R. Socher 2019. Ctrl: A Conditional Transformer Language Model For Controllable Generation. arXiv preprint arXiv:1909.05858.

Khoubrouy, S. A., and J. H. Hansen. 2016. "Microphone Array Processing Strategies for Distant-Based Automatic Speech Recognition." *IEEE Signal Processing Letters* 23(10): 1344–8.

Khouzaimi, H., R. Laroche, and F. Lefèvre. 2018. "A Methodology for Turn-Taking Capabilities Enhancement in Spoken Dialogue Systems Using Reinforcement Learning." *Computer Speech and Language* 47: 93–111.

Ko, T., V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur. 2017. "A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5220–4). IEEE.

Kocielnik, R., D. Avrahami, J. Marlow, D. Lu, and G. Hsieh. 2018. "Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent." In Proceedings of the 2018 Designing Interactive Systems Conference (pp. 881–94). ACM.

Kong-Vega, N., M. Shen, M. Wang, and L. F. D'Haro. 2019. "Subjective Annotation and Evaluation of Three Different Chatbots WOCHAT: Shared Task Report." In *9th International Workshop on Spoken Dialogue System Technology* (pp. 371–8). Singapore: Springer.

Kostelník, P., I. Pisařovic, M. Muroň, F. Dařena, and D. Procházka. 2019. "Chatbots for Enterprises: Outlook." *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis* 67(6): 1541–50.

Kurachi, Y., S. Narukawa, and H. Hara. 2018. "AI Chatbot to Realize Sophistication of Customer Contact Points." *Fujitsu Scientific and Technical Journal* 54: 2–8.

Li, S., and C. H. Chen. 2019. "The Effects of Visual Feedback Designs on Long Wait Time of Mobile Application User Interface." *Interacting with Computers* 31(1): 1–12.

Li, K., H. Xu, Y. Wang, D. Povey, and S. Khudanpur. 2018. "Recurrent Neural Network Language Model Adaptation for Conversational Speech Recognition." *Proc. Interspeech* 2018, 3373–7, DOI: 10.21437/Interspeech.2018-1413.

Li, X. B., and J. Qin. 2017. "Anonymizing and Sharing Medical Text Records." *Information Systems Research* 28(2): 332–52.

Lister, K., T. Coughlan, F. Iniesto, N. Freear, and P. Devine. 2020. "Accessible Conversational User Interfaces: Considerations for

Design." In Proceedings of the 17th International Web for All Conference (pp. 1–11).

Litman, D. J., M. A. Walker, and M. S. Kearns. 1999. "Automatic Detection of Poor Speech Recognition at the Dialogue Level." In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 309–16). Association for Computational Linguistics.

Liu, X., A. Eshghi, P. Swietojanski, and V. Rieser. 2019a. Benchmarking Natural Language Understanding Services for Building Conversational Agents. In: Marchi E., S. M. Siniscalchi, S. Cumani, V. M. Salerno, and H. Li, eds. Increasing Naturalness and Flexibility in Spoken Dialogue Interaction. Lecture Notes in Electrical Engineering, vol. 714. Springer, Singapore. https://doi.org/10.1007/978-981-15-9323-9_15

Liu, H., J. Dacon, W. Fan, H. Liu, Z. Liu, and J. Tang. 2019b. Does Gender Matter? Towards Fairness in Dialogue Systems. Proceedings of the 28th International Conference on Computational Linguistics (COLING), pp. 4403–16

López-Cózar, R., Z. Callejas, D. Griol, and J. F. Quesada. 2014. "Review of Spoken Dialogue Systems." *Loquens* 1(2): e012.

Mani, A., S. Palaskar, N. V. Meripo, S. Konam, and F. Metze. 2020. "ASR Error Correction and Domain Adaptation Using Machine Translation." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP (pp. 6344–8). IEEE.

Mehri, S., and M. Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 681–707.

McTear, M. F., Z. Callejas, and D. Griol. 2016. *The Conversational Interface* (Vol. 6, No. 94, p. 102). Cham: Springer.

Nuruzzaman, M., and O. K. Hussain. 2018. "A survey on Chatbot Implementation in Customer Service Industry Through Deep Neural Networks." In 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE) (pp. 54–61). IEEE.

Ondáš, S., J. Juhár, E. Kiktová, and J. Zimmermann. 2018. "Anticipation in Speech-based Human-Machine Interfaces." In 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom) (pp. 000117–22). IEEE.

Park, D. S., W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le 2019. Specaugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Proc. Interspeech 2019, 2613–17, DOI: 10.21437/Interspeech.2019-2680.

Polino, A., R. Pascanu, and D. Alistarh 2018. Model Compression Via Distillation And Quantization. Sixth International Conference on Learning Representations (ICLR).

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." *OpenAI Blog* 1(8): 9.

Ramesh, K., S. Ravishankaran, A. Joshi, and K. Chandrasekaran. 2017. "A Survey of Design Techniques for Conversational Agents." In International Conference on Information, Communication and Computing Technology (pp. 336–50). Springer, Singapore.

Ray, A., and R. Mathew. 2018. "Review of Cloud-Based Natural Language Processing Services and Tools for Chatbots." In International conference on Computer Networks, Big data and IoT (pp. 156–62). Springer, Cham.

Reddy, T. 2017. The Code of Ethics for AI And Chatbots That Every Brand Should Follow. IBM. Available at https://www.ibm.com/blogs/watson/2017/10/the-code-of-ethics-for-ai-and-chatbots-that-every-brand-should-follow/

Ruan, S., J. O. Wobbrock, K. Liou, A. Ng, and J. A. Landay. 2018. "Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1(4): 1–23.

Ruane, E., A. Birhane, and A. Ventresque. 2019. Conversational AI: Social and Ethical Considerations.

Ruder, S., M. E. Peters, S. Swayamdipta, and T. Wolf. 2019. "Transfer Learning in Natural Language Processing." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials (pp. 15–18).

Saglam, R. B., and J. R. Nurse 2020. Is Your Chatbot GDPR Compliant? Open Issues In Agent Design. CUI '20: Proceedings of the 2nd Conference on Conversational User Interfaces, July 2020 Article No. 16, pp. 1–3.

Samanta, S., and S. Mehta 2017. Towards Crafting Text Adversarial Samples. arXiv preprint arXiv:1707.02812.

Sarkar, D. 2018. A Practitioner's Guide to Natural Language Processing (Part I) — Processing and Understanding Text. Towards Data Science (blog). Available at https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72

Seki, H., T. Hori, S. Watanabe, J. L. Roux, and J. R. Hershey 2018. A Purely End-To-End System for Multi- Speaker Speech Recognition. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2620–30.

Schultz, T. and K. Kirchhoff, eds. 2006. *Multilingual Speech Processing*. 1. Elsevier (pp. 1–536).

See, A., S. Roller, D. Kiela, and J. Weston 2019. What Makes a Good Conversation? How Controllable Attributes Affect Human Judgments. NAACL-HLT (1) 2019: 1702–23.

Serban, I. V., R. Lowe, P. Henderson, L. Charlin, and J. Pineau. 2015. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. CoRR abs/1512.05742.

Serizel, R., and D. Giuliani. 2017. "Deep-neural Network Approaches for Speech Recognition with Heterogeneous Groups of Speakers Including Children." *Natural Language Engineering* 23(3): 325–50.

Shafey, L. E., H. Soltau, and I. Shafran. 2019. Joint Speech Recognition and Speaker Diarization Via Sequence Transduction. Proc. Interspeech 2019, 396–400, DOI: 10.21437/Interspeech.2019-1943.

Shanbhag, A. 2020. 5 Chatbot Code of Ethics Every Business Should Follow. Available at https://botcore.ai/blog/5-chatbot-code-of-ethics-every-business-should-follow/

Smestad, T. L. 2018. Personality Matters! Improving the User Experience of Chatbot Interfaces- Personality Provides a Stable Pattern to Guide the Design and Behaviour of Conversational Agents (Master's Thesis, NTNU).

Sproat, R., and N. Jaitly 2016. RNN Approaches to Text Normalization: A challenge. CoRR abs/1611.00068.

Song, Y., R. Yan, C. T. Li, J. Y. Nie, M. Zhang, and D. Zhao 2018. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), pp. 4382–8.

Tammewar, A., M. Pamecha, C. Jain, A. Nagvenkar, and K. Modi. 2018. "Production Ready Chatbots: Generate if not Retrieve." In

Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.

Tao, C., L. Mou, D. Zhao, and R. Yan. 2018. "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems." In Thirty-Second AAAI Conference on Artificial Intelligence.

Vanian, J., and A. Pressman. 2019. "How Amazon, Apple, Google, and Microsoft Created an Eavesdropping Explosion—Data Sheet." *Fortune*. Available at https://fortune.com/2019/08/08/gogole-amazon-microsoft-listen-conversation-siri/

Vasserman, L., B. Haynor, and P. Aleksic. 2016. "Contextual Language Model Adaptation Using Dynamic Classes." In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 441–6). IEEE.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Wieting, J., and K. Gimpel. 2017. Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. arXiv preprint arXiv:1711.05732.

Wolf, M. J., K. Miller, and F. S. Grodzinsky. 2017. "Why We Should Have Seen That Coming: Comments on Microsoft's Tay Experiment, and Wider Implications." *ACM SIGCAS Computers and Society* 47(3): 54–64.

Worswick, S. 2019. Mitsuku Wins Loebner Prize 2019 and Best Overall Chatbto at AISB X. Available at https://aisb.org.uk/mitsuku-wins-2019-loebner-prize-and-best-overall-chatbot-at-aisb-x/

Yuan, Y., S. Thompson, K. Watson, A. Chase, A. Senthilkumar, A. B. Brush, and S. Yarosh. 2019. "Speech Interface Reformulations and Voice Assistant Personification Preferences of Children and Parents." *International Journal of Child-Computer Interaction* 21: 77–88.

Young, S. 2002. The Statistical Approach to the Design of Spoken Dialogue Systems. Tech Report CUED/F-INFENG/TR.433, Cambridge University Engineering Department.

Yuwono, S. K., B. Wu, and L. F. D'Haro. 2019. "Automated Scoring of Chatbot Responses in Conversational Dialogue." In 9th International Workshop on Spoken Dialogue System Technology (pp. 357–69). Springer, Singapore.

Zhang, C., L. F. D'Haro, R. E. Banchs, T. Friedrichs, and H. Li, 2020. *Deep AM-FM: Toolkit for Automatic Dialogue Evaluation*. Springer.

AUTHOR BIOGRAPHIES

Javier Cebrián is a project manager at Saturno Labs, he has participated in the development of chatbots for commercial applications. He has a Bachelor's degree in Telecommunication Technologies and Services Engineering from Universidad Politécnica de Madrid (ETSIT, UPM) and collaborates as a researcher with the Speech Technology Group (GTH@UPM) in the development of applications for studying sleep apnea. Currently, he is studying a Masters in Telecommunication Engineering at the Universitat Oberta de Catalunya (UOC).

Ramón Martínez is a project manager at Saturno Labs, he has participated in the development of chatbots for commercial applications. He has a Bachelor's degree in Telecommunication Technologies and Services Engineering from Universidad Politécnica de Madrid (ETSIT, UPM) in Spain. Currently, he is studying Master's degree in Signal Processing and Machine Learning for Big Data at the Universidad Politécnica de Madrid (ETSIT, UPM).

Natalia Rodríguez is the founder and CEO at Saturno Labs, an innovation lab. She has a Bachelor's Degree in Telecommunication Technologies and Services Engineering, and currently, she is studying a Master's degree in Signal Processing and Machine Learning for Big Data at the Universidad Politécnica de Madrid (ETSIT, UPM). Her current research and work mainly focuses on product engineering and business development. Her concepts have been used by thousands of users around the world. She has created more than 15 technological products (either by herself or leading groups of programmers) among which are: mobile apps for multinationals in the retail world, banking or press, platforms for schools, institutions and companies, several chatbots and Alexa skills and machine and Deep Learning products for data analysis. In addition, she has created a social network with presence in 27 countries, she has led a technical project in 62 countries in a multinational corporation and she has been a counselor in the "Special Advisory Group" to organize an event similar to the World Economic Forum on education and employment. In addition, she is a professor of project management, project generation and application of the latest technologies to business development at Sae Institute. She has won more than 15 individual awards and hackathons for her work.

Luis Fernando D'Haro is an Associate Professor at the Universidad Politécnica de Madrid (ETSIT, UPM) in Spain and a member of the Speech Technology Group (Speech@UPM). His current research mainly focuses on spoken dialogue and natural language processing systems; he has written more than 15 international publications specifically on dialog systems. He co-led the International Dialog State Tracking Challenges (DSTC) in 2015 and 2016 and is a member of the program and steering committees from DSTC6 to DSTC9 challenges. Since 2015, Prof. D'Haro has co- organized the WoChat series of workshops, which have the common goal of advancing chatbot systems and their automatic evaluation. He was also member of the local organizers for Interspeech in 2014, Human Agent Interaction conference (HAI2016), and the International Workshop on Spoken Dialog System Technology (IWSDS) in 2018; he is currently the general

chair for IWSDS 2020 to be held in Madrid, Spain. Luis Fernando is also working on an initiative for collecting and annotating data from human-chatbots interactions called WOCHAT with the goal of designing new algorithms to automatically evaluate dialogue systems, as well as senior member for the Johns Hopkins Summer school (JSALT2020).

How to cite this article: Cebrián, J., R. Martínez, N. Rodríguez, and L. F. D'Haro. 2021. "Considerations on Creating Conversational Agents for Multiple Environments and Users." *AI Magazine* 42: 71–86. https://doi.org/10.1609/aaai.12007.