CORK: A COnversational agent framewoRK exploiting both rational and emotional intelligence

Conference Paper · March 2019

CITATIONS
12

READS
257

4 authors:

Fabio Catania
Politecnico di Milano
22 PUBLICATIONS **145** CITATIONS
SEE PROFILE

Micol Spitale
University of Cambridge
38 PUBLICATIONS **163** CITATIONS
SEE PROFILE

Davide Fisicaro
Politecnico di Milano
4 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Franca Garzotto
Politecnico di Milano
275 PUBLICATIONS **5,268** CITATIONS
SEE PROFILE

CORK: A COnversational agent framewoRK exploiting both rational and emotional intelligence

Fabio Catania
fabio.catania@polimi.it
Politecnico di Milano
Milano, Italy

Micol Spitale
micol.spitale@polimi.it
Politecnico di Milano
Milano, Italy

Davide Fisicaro
davide.fisicaro@polimi.it
Politecnico di Milano
Milano, Italy

Franca Garzotto
franca.garzotto@polimi.it
Politecnico di Milano
Milano, Italy

ABSTRACT

This paper proposes CORK, a modular framework to facilitate and accelerate the realization and the maintenance of intelligent Conversational Agents with both rational and emotional capabilities. A smart CA can be integrated in any digital device and aims at interpreting the natural language inputs by the user and at responding to them consistently with the semantics, the context, the user's perceived emotional state and her/his profile.
CORK's strength is its ability to split the content of the speech from its pure conversational scheme and rules: it permits to the developers to detect during design some general communicative patterns to be filled at run-time with the right content depending on the context. At first, this can be tiresome and time-consuming, but it permits then to achieve high scalability and low-cost modification and maintainability.
This framework is potentially valid in general and for now it has been effectively applied to develop conversational tools for people with neurodevelopmental disorders.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; *Empirical studies in HCI*.

KEYWORDS

Conversational technology; Affective computing; Conversational Agent Framework;

ACM Reference Format:
Fabio Catania, Micol Spitale, Davide Fisicaro, and Franca Garzotto. 2019. CORK: A COnversational agent framewoRK exploiting both rational and emotional intelligence. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019* , 8 pages.

1 INTRODUCTION

A Conversational Agent (CA), or dialogue system, is a software program able to interact through natural language. Voice-based CAs and chatbots are progressively getting more and more embedded into people's home and daily routines: the most famous are Apple's Siri, Amazon's Alexa (over 8 million users in January 2017), Google Assistant and IBM conversational services.
In usual human-human interaction, non-verbal information is responsible for about 93% of the message perception [7] and its relevance has been recently investigated a lot in human computer interfaces, too. Affective computing is the interdisciplinary research field regarding systems and devices that can recognize, interpret, process and simulate human affects. Currently, it is one of the most active research topics, furthermore, having increasingly intensive attention. According to BusinessWire [5], global Affective Computing market has been valued at USD 16.17 billion in 2017 and is expected to reach a value of USD 88.69 billion by 2023.
Emotional Conversational Agents are at the very beginning and still no generally used architectural framework has been proposed in order to facilitate their implementation and fair comparison.
Conversational Technology's designers know how tiring and time-consuming is the process to create a dialog system and to maintain it: it is necessary to detect and anticipate the inordinate ways that a user may send input to the system and to couple them with the best outputs. What is more is that modifying the content of the output implies to operate directly on the conversation. The ideal Conversational Agent should be able to adapt to different situations and scenarios and to be easily correctable in terms of content.
In this respect, this work provides a clear contribution to advance the state-of-the art: we propose CORK, a modular framework to facilitate and accelerate the realization and the maintenance of intelligent, system-initiative Conversational Agents with both rational and emotional capabilities. A big and innovative challenge faced by CORK is the split of the content of the speech from its pure conversational scheme and rules. This permits to detect during design some general communicative patterns to be filled at run-time with different contents depending on the context.
This framework is potentially valid in general and for now it has been effectively applied to develop conversational tools for people with neurodevelopmental disorders NDD. In particular, we investigate the use of Emoty, a spoken Conversational Agent realized with our framework, to mitigate the impairments of these persons related to the difficulty of recognizing and expressing emotions - a problem clinically referred to as Alexithymia.

IUI Workshops'19, March 20, 2019, Los Angeles, USA

2 STATE OF THE ART

ELIZA [39], one of the earliest developed CAs, was capable of creating the illusion that the system was actually listening to the user without even using any built-in framework for contextualizing events by using a simple pattern matching and substitution methodology.
As technology has evolved, retrieval-based conversational agents have become the most used ones. They get as input an assertion and a context (the conversation up to now); then, they use some heuristic functions to pick up an appropriate response from a predefined repository. The heuristic could be either a complex group of Machine Learning (ML) classifiers or a simpler rule-based expression match. This way, the set of possible responses is fixed, but at least is grammatically correct. The chatbots realized with the help of Dialogflow [12], wit.ai [11], Alexa [2], Watson [19] and Azure bot service [24] are all examples of rationally-intelligent retrieval-based conversational agents. The main disadvantage of this model is that potentially large numbers of rules and patterns would be required. Rule elicitation is undoubtedly a time-consuming, high maintenance task. Secondly, modifying one rule or introducing a new one into the script invariably has an impact on the remaining rules [26].
On the other hand, generative models do not rely on pre-defined responses but generate new answers from scratch using some ML techniques. Unfortunately, nowadays they do not work well, yet, because of the fact that they tend to make grammatical mistakes and to produce irrelevant, generic or inconsistent responses. In addition, they need a huge amount of training data and are very hard to optimize. Generative models represent an active area of research [22], [33]. Much research is being focused on pursuing socially and emotionally aware CAs [4], [27].
Human-computer interaction can be more effective taking into account the emotional state of the user. Cassell et al. [6] cites that embodied CAs should now be capable of developing deeper relationships that make their collaboration with humans productive and satisfying over long periods of time. Unfortunately, the basic versions of both retrieval-based and generative models do not take into account any emotional aspect within the conversation.
We studied the architecture of some affective conversational agents, e.g. [16], [34], and we found out that all of them use an upgrade of the retrieval-based model exploiting the result of emotional analysis as an element of choice for the final answers. We were surprised to note that in literature there is no generally used, shared framework for the realization of emotionally aware dialog systems.

3 THEORETICAL FUNDAMENTALS

According to the American theorist and university professor David Berlo [10] there exist four main factors in the human communication process:

- the Source, who is the sender of the message;
- the Message, that includes the content of the communication, its structure, the context and the form the message is sent;
- the Channel, that is the medium used to send the message (looking, listening, tasting, touching and smelling);
- the Receiver, who is the person who gets the message and tries to understand what the sender wants to convey in order to respond accordingly.

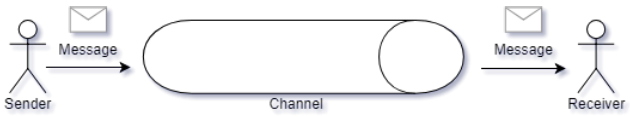

Figure 1: Berlo's communication model

In this paper, we promote the application of these concepts to conversational technologies and we assert that the basis of Berlo's theory can be adapted to the human-computer natural language interaction, where the roles of the sender and the receiver are played alternately by a human being and an intelligent machine.
Let's think about a system initiative Conversational Agent, that is a dialog system that needs to completely control the flow of the conversation asking to the user a series of questions and ignoring (or misinterpreting) anything the user says that is not a direct answer to the system's question. In this setting, every time the system transmits a message to the user, it attaches also the context it refers to. When the user answers, the message and the same context are sent back to the system. At this point, the dialog system decodes and elaborates the user's request (message plus context) and replies consistently back.

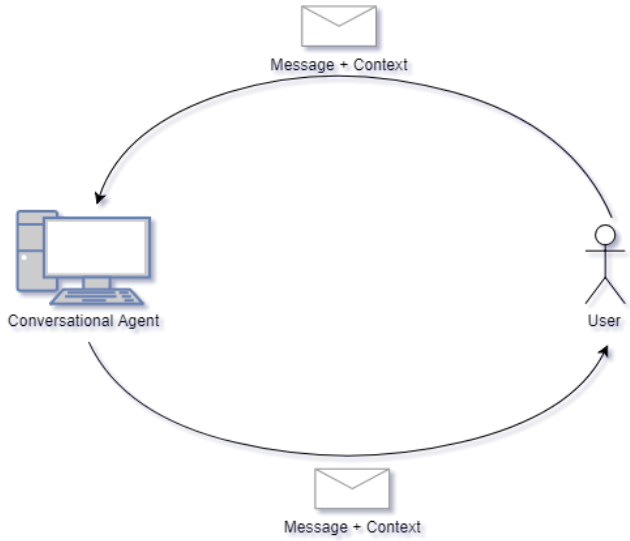

Figure 2: Adaptation of Berlo's model to Conversational Technology

The exploited channel depends from the nature of the conversational agent. Chatbots use written text as a conduit, while spoken dialogue systems use the verbal channel and have two essential components: a speech recognizer and a text-to-speech module. In

addition, Conversational Agents can take advantage of the visual channel thanks to the use of a camera and can be embodied by avatars and robots and exploit the physical channel to understand and communicate to the user.

4 THE FRAMEWORK

In order to hold a conversation, a dialog system must be able to finalize three main phases:

- the decoding and understanding of the natural language input thanks to the input recognizer;
- the execution of a consistent task and/or the generation of a logical output;
- the rendering of the output.

Designing this framework, we reason a Spoken Conversational Agent with emotional sensitivity. We propose a client-server software architecture exploiting the Model-View-Controller MVC architectural pattern. As is well known, it divides the software application into three interconnected logical tiers, so that it separates the internal representations of information from how information is elaborated and from the ways that it is presented to the user. This choice is to lend flexibility, robustness and scalability to the system. In addition, this way, it will be possible to easily integrate the realized CAs in digital devices such as tablets or smart phones or embed them in everyday physical objects (e.g. toys, home equipment).
Each tier is described as follows:

- client Tier: it is the topmost level of the application and is thought as thin as possible. Its only goal is to manage the inputs and the outputs by the user during the whole conversational session. This means that it records the user while speaking, it sends it to the server, it waits for an answer and produces an artificial human speech as an output;
- logic Tier: it processes the input message coming from the client tier with the help of external services and it interacts with the persistence tier to select the best possible answer for the user;
- persistence Tier: it is in charge of storing and retrieving information from the database (i.e., data about the users, the preset output templates, the conversation contents, ...).

As a matter of higher reusability and maintainability, we propose a modular software architecture with independent blocks, such that each contains everything necessary to execute only one aspect of the main elaboration. The modules constituting the early version of our architecture are the following:

- the recorder;
- the engine;
- the speech-to-text module;
- the NLP Unit and intent detector;
- the sentiment analysis module;
- the emotional analysis module;
- the topic analysis module;
- the voice analysis module;
- the profiling module;
- the output creation module;
- the text-to-speech module.

Other modules executing additional elaborations can be integrated in a later stage.

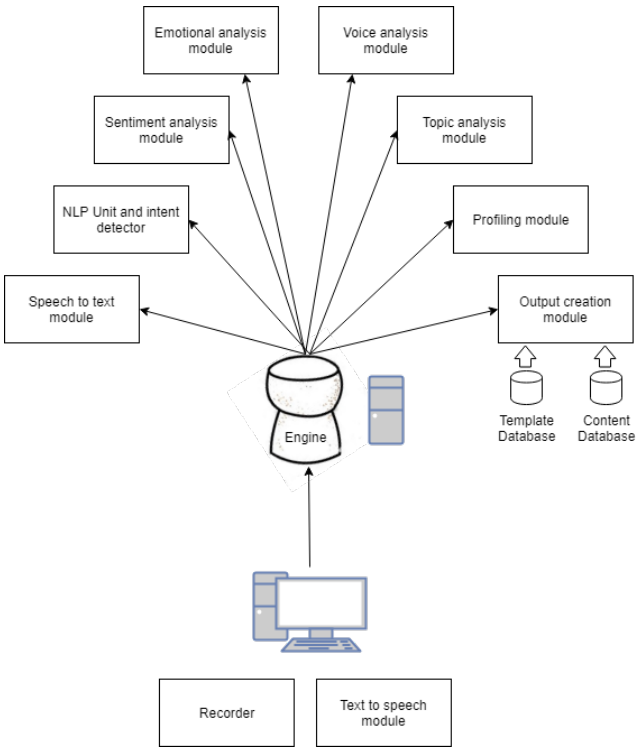

Figure 3: Functional view of the modules composing CORK

Below you can find a detailed description of each component.

4.1 The recorder

This module lies on the client and, as its name suggests, is responsible for recording the user when she/he speaks and tries to interact with the system. Beyond that, it is in charge of sending to the server, in particular to the engine module, the recorded audio file and the conversational context referred to. The main challenges faced by the recorder module are the identification of the instants when to start and to complete the recording (that ideally correspond to the moments when the user starts and finish to speak). Optionally, cutting-edge versions of this software component can run an early audio data processing to clean it from background noises and the speaker diarisation. Diarisation is the process of partitioning an input audio stream into homogeneous segments according to the speaker identity.

4.2 The engine

The logic and the control of the whole system lie in the engine module. At every conversational step, this receives the message and the relative context from the client tier; depending on the exploited channel, the message can be an audio recording, a written text, a picture or a rap. Successively, the engine interprets and elaborates the message by calling one by one the other components in order

to generate the best answer for the user; the processing may vary with respect to the context and the nature of the input message. For example, some modules may be excluded from the process and some elaborations may be parallelized. A basic elaboration flow is represented by Figure 4. Finally, the generated output message and context are sent back to the client tier.

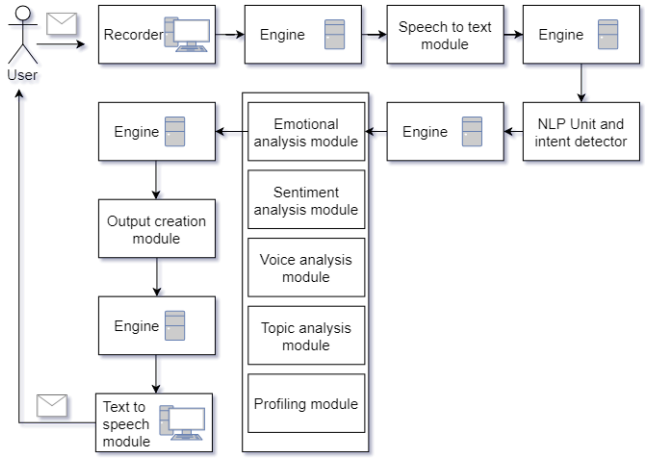

Figure 4: The message processing flow

4.3 The speech to text module

The speech to text module gets as input a prerecorded audio containing some spoken words. It applies advanced deep-learning and machine learning neural network algorithms to recognize and transcript the speech into text and return it as a string. Some speech recognition systems require "training" where an individual speaker reads text into the system. In this way, the system analyzes then the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Especially because of the large amount of data they have at their disposal to train their models, the cutting-edge speech to text services are Google Cloud Speech-to-Text [13], IBM Watson Speech to Text [18] and Microsoft Azure Speech to Text [25]. They work for a wide range of languages and are easily embeddable via API.

4.4 The NLP unit and intent detector

The Natural Language Processing Unit and Intent Detector is one of the most relevant modules of the whole system. It receives as input a string of text representing the transcription of the user's speech and the context it refers to. By exploiting domain knowledge and natural language comprehension capabilities, it analyzes, understands and returns the user's intent. Intents are links between what the user says and its interpretation by the bot. Contexts are useful for differentiating requests which might have different meaning depending on previous requests. The platforms Dialogflow [12], wit.ai [11], Alexa [2], Watson [19] and Azure bot service [24] use an intent-based approach: that means that they elaborate a reaction to the user's intention detected by the NLP unit. Again, the large amount of data they have at their disposal to train the models, makes their NLP units the state-of-the-art technologies to detect the intention of the user in every conversation step.

4.5 The sentiment analysis module

Sentiment analysis regards finding out whether a given text expresses positivity, neutrality or negativity. This module gets as input the speech by the user and returns its polarity back. The most advanced APIs on the marketplace are Google cloud - natural language [15], Azure - text analytics [23], Repustate [31], Text processing [29] and TextRazor [35].

4.6 The emotional analysis module

Emotion analysis permits to extract the feelings expressed by a text using Natural Language Processing. Feelings are in this case the Big Six emotions (Joy, Sadness, Fear, Anger, Surprise, Disgust) and other cognitive statements as confidence, uncertainty and attention. This module gets as input the transcription of the speech by the user. After its examination, the service returns a dictionary that maps the emotions and cognitive statements to the probability that the author is expressing them. The most advanced APIs of this kind are IBM Tone Analyzer [17], Indico.io [21], Q^{o}Emotion [30] and TheySay [36] and they work all with English language.

4.7 The topic analysis module

The topic analysis module uses machine learning and natural language processing to discover the abstract topics that occur in a written work. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a written text is about a particular topic, one would expect particular words to appear in the text more or less frequently. At the moment of writing, Google Cloud Natural Language [15], TextRazor [35] and Meaning Cloud Topic Extraction [8] are three relevant and advanced API providing this kind of service.

4.8 The profiling module

The profiling module customizes the contents and the style of the conversation depending on every single user. More in detail it exploits relevant info about the user (generalities, preferences, needs, ...) to influence the best output to be provided to the user and the best way to do it. To do so, it interacts with the persistence tier: it is in charge of storing the info and of picking them up if necessary. Relevant info can be obtained during previous conversations, providing the dialog systems with episodic memory. Alternatively, they can be received through other channels, such as social networks, pre-configuration by the user, third party configuration.

4.9 The voice analysis module

This module is in charge of understanding the emotion in a speaker's voice. It gets as input an audio recording containing a human speech and returns back the emotion perceived analyzing the pitch and the intonation of the voice. Emotional speech analysis starting from the harmonic features of the audio is a relatively new research field and there is much room for improvement. The most relevant APIs providing this kind of service are Good Vibrations [37], Affectiva [1] and Vokaturi [38].

4.10 The output creation module

This innovative module permits the customization of the output to be delivered to the user in order to obtain a user-centered content driven conversational agent. In a first stage, it picks up the best fitting answer from a table of anticipated templates according to

- the state of the conversation,
- the detected intention of the user,
- the user in case,
- the sentiment detected from the semantic,
- the emotion recognized both from the text and from the pitch of the voice of the user's message.

In a second stage, it fills the selected template with the right content chosen from the contents table according to the result of the topic analysis and, again, to the user's intention. The final generated output is returned to the engine module. This kind of table-driven architecture permits the easy update and modification of the contents without even touching the conversational structure.

4.11 The text to speech module

The text to speech module is responsible for the human voice synthesis, that is the artificial production of human speech. Compared to recorded human speech, the advantage of synthetized voice is that its content can change and be customized at runtime. By playing around with the voice features, the text to speech module may express emotions, moods and cognitive states. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood clearly. As for the speech to text module, some speech synthesizers permit "training" where an individual speaker reads text into the system. In this way, the system analyzes then the person's specific voice and uses it to fine-tune the duplicate of that person's typical sounds. There exists many text to speech APIs available on the market, such as Google Cloud Text to Speech [14], Amazon Polly [3], IBM Watson Text to Speech [20] and ResponsiveVoice [32]. They all provide female and male voices speaking different languages and with different accents.

5 DEVELOPED CONVERSATIONAL AGENTS

The CORK framework is potentially valid in general. For now, it has been applied to develop conversational tools for people with neurodevelopmental disorders NDD. NDD is a group of conditions that are characterized by severe deficits in the cognitive, emotional and motor areas and produce impairments of social functioning. Any kind of NDD is a chronical state, but early and focused interventions are thought to mitigate its effect. As an experiment, we investigated the use of conversational technologies as a tool for persons with NDD to lead them to enhance their emotion expression and communication capabilities and to get more socially integrated.

5.1 Ele

As a first part of our work, we developed Ele (see Figure 5), an embodied Wizard-of-Oz dialog system with the aspect of an elephant. This robot wants to be a conversational companion that speaks through the live voice of a remote human operator and can engage the user in dialogues and recounting stories, enriching the communication using toys body movement effects. In return, the user can interact with the social robot verbally, making facial expressions and by touch. This system has been used to detect some relevant conversational patterns for people with NDD: for up to one month we observed with the help of an expert the weekly sessions of a group of 11 NDD adults aged 25 to 43 interacting with Ele. From this experience and three weekly meeting sessions with two psychological specialists it turns out that some ways of speaking (with a lot of repetitions, reassurances and continuous reinforcements) are the most effective for people with NDD.

Figure 5: Ele

5.2 Emoty

Once we detected a considerable number of conversational patterns for people with neurodevelopmental disorders, we designed a tool for people affected by a specific disturb called Alexithymia, that is the inability to identify and express emotions. In particular, we developed Emoty, a Dialog System playing the role of emotional facilitator and trainer by exploiting emotion detection capabilities from both text and audio data. Emoty proposes to the user some emotion expression tasks in the form of a game, organized in increasingly difficult levels. For example, the user could be asked to read an assigned sentence trying to express a given emotion with her/his tone of voice.

Figure 6: Emoty

The project has been designed in close collaboration with caregivers and psychologists who actively participated in the following phases:

(1) eliciting the key requirements,
(2) evaluating iterative prototypes,
(3) performing an exploratory evaluation.

In addition to the verbal channel, the application exploits the visual one as a support to the player (see Figure 6). Specifically, the use of emojis facilitates the conceptualization of the emotions, and the combination emotion-color works as reinforcement to the spoken feedbacks by the system. This kind of matching is very common in literature and we decided to follow Plutchik's theory [28]. In his model, joy is represented as yellow, fear is dark green, surprise is cyan, sadness is blue, and anger is red. We associated neutrality to light grey, used as background for the system as well.

5.2.1 Architectural overview. The system follows the guidance of CORK, the Conversational agent framework. It has been realized as a web application because web apps are accessible to everybody via browser on everyday devices and do not need any kind of preliminary configuration. They permit both vocal and visual interaction with the user thanks to the screen, the microphone and the speakers of the executor device.
The engine module of the entire system lies on the Cloud and is accessible through serverless functions to be triggered via HTTPS requests. This guarantees fair pricing, a safe execution environment and high-level scalability and availability.
For every conversation step, the engine remotely calls Google Cloud Speech-to-Text to get a transcription of the user's speech and then Dialogflow to detect the user's intention according to the context. The semantic analysis of the input contents is delegated by our system to indico.io. This external service returns a dictionary that maps anger, fear, joy, sadness and surprise to the probability with which the author is expressing each emotion. At the time of development, there is no emotion analysis API working in Italian language and this constrains us to translate our Italian inputs to English and then to let them be processed by the emotion recognizer. The automatic translation may imply a loss of information to be considered, however after a preliminary evaluation the results are satisfactory.
The emotion recognition from the harmonic features of the audio is performed by an original machine learning model all by us. The process is organized in two subsequent steps:

(1) feature extraction
(2) classification into emotions.

The former extracts temporal and spectral characteristics of the audio signal; whereas, the classification has been implemented using a supervised learning approach based on deep neural networks. In particular, a wide and deep (convolutional) neural network has been built to exploit the principle of temporal locality across audio signal partitions and increase the discriminatory strength of the model. In order to properly train the model an open source and free Italian dataset called Emovo [9] has been used. It is a corpus of recordings by 6 actors playing 14 different sentences simulating the five emotion categories.
The messages by the system are generated by the output creation module that fills the conversational templates detected with Ele with the right contents depending on the context and on the interacting user. Information about the users are inserted before the first session by the caregiver supporting her/him. They are the name, the age and some parameters indicating the ability of the user to express and recognize emotions.
Finally, the human voice synthesis is done by calling the Google Cloud Text to Speech API.

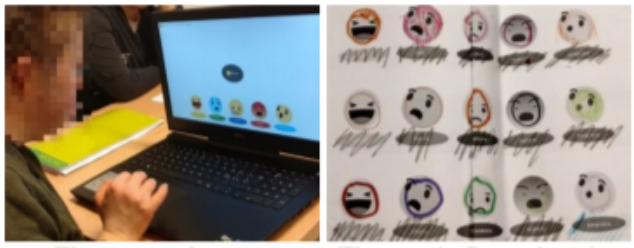

Figure 7: A user playing with the application and a drawing by a user as a side activity

5.2.2 Exploratory study. When ready, Emoty has been field tested by people with NDD at mild or moderate severity ranging from 16 to 60 years old. Our study has been plan to take place in a six-month evaluation organized in weekly scheduled sessions of 10-15 minutes integrated in daily activities at a specialized care center. Experimental sessions are focusing on both

- the improvements in performance (i.e., the completion of the activities) of users across sessions in time;
- the usability of the application.

For now, only three weeks of experimentation have been conducted. Already after this little time, according to the attending therapist some users tend to be more eager to interact with our device rather than to speak to other people. We know that the evaluation based on three sessions is hardly statistically significant, therefore, for now, the caregivers' opinions are the most valuable feedback we can collect: they observed a growing awareness about feelings by the participants. Anyway, the graphical representation of the performances by three users across the three sessions is reported in Figure 8.

This encouraging early results are a clue that the detected conversational patterns fit the target user's needs and the conversational agent framework was well designed to develop complex emotionally sensitive dialog systems (at least for this application domain).

6 CONCLUSION

This work proposes a solution to a common need we observed in literature: the necessity to have a structured, tested and simplified way to develop a conversational agent.
In this paper we described CORK, a modular framework to facilitate and accelerate the realization and the maintenance of intelligent Conversational Agents with both rational and emotional capabilities.

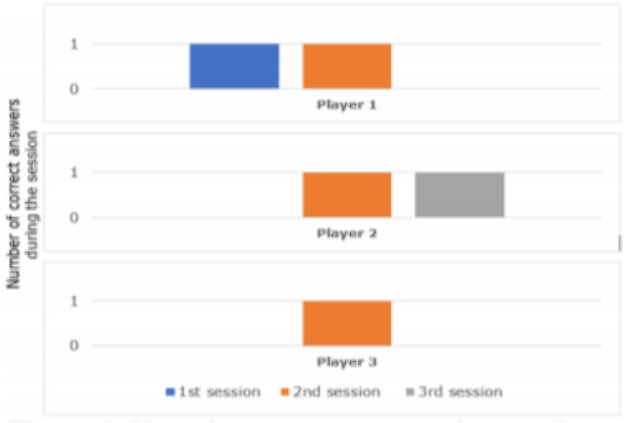

Figure 8: Users' progress across the sessions playing with Emoty

Our framework has a client-server architecture exploiting the Model-View-Controller MVC pattern. This way, it will be possible to easily integrate the realized CAs in digital devices such as tablets or smart phones or embed them in everyday physical objects (e.g. toys, home equipment). In addition, the system is organized in independent, reusable software modules controlled by a centralized engine, such that each of them deals with a single functionality in the whole conversational process. The proposed components are the recorder, the engine, the speech-to-text module, the NLP Unit and intent detector, the sentiment analysis module, the emotional analysis module, the topic analysis module, the profiling module, the output creation module and the text-to-speech module. Other modules executing additional elaborations can be integrated at will.

A big and innovative challenge faced by CORK is the split of the content of the speech from its pure conversational scheme and rules. This permits to detect during design some general communicative patterns to be filled at run-time with different contents depending on the context.

To test the effectiveness of the proposed framework, we designed and developed in cooperation with psychologists and therapists an emotional facilitator and trainer, called Emoty, for individuals with neurodevelopmental disorders as a conversational supporting tool for regular interventions. Although Emoty is still a prototype, the initial results of our empirical study indicate that Emoty can be easily used by therapists and persons with NDD and has some potential to mitigate Alexithymia and its effects. After only three weeks of practice with the emotional trainer, the caregivers observed a growing awareness by the users of their own feelings. This encouraging early results are a clue that the chosen conversational patterns fit the target user's needs and the conversational agent framework was well designed to develop complex emotionally sensitive dialog systems for this application domain and in general at large.

The natural follow up of this project will be the development of a user-friendly platform inspired to IBM Watson Assistant and Dialogflow that permits the development of content-driven emotionally sensitive Conversational Agents without even programming. In addition, we will try to improve the effectiveness of the framework by testing it with new dialog systems in new application domains. In parallel, we will work to improve the accuracy of the emotion recognition Machine Learning ML model. To do so, we will create and tag a larger, proprietary emotional dataset collecting speech-based conversations and dialogues with the collaboration of local theatre companies and acting schools.

REFERENCES

[1] Affectiva. 2018. Affectiva. https://www.affectiva.com
[2] Amazon. 2018. Alexa. https://developer.amazon.com/it/alexa
[3] Amazon. 2018. Amazon Polly. https://aws.amazon.com/polly/?nc1=f_ls
[4] Christian Becker, Stefan Kopp, and Ipke Wachsmuth. 2007. Why emotions should be integrated into conversational agents. *Conversational informatics: an engineering approach* (2007), 49–68.
[5] BusinessWire. 2018. Global Affective Computing Market 2018-2023. https://bit.ly/2Qv5kKX
[6] Justine Cassell, Alastair J Gill, and Paul A Tepper. 2007. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing*. Association for Computational Linguistics, 41–50.
[7] Claude C Chibelushi and Fabrice Bourel. 2003. Facial expression recognition: A brief tutorial overview. *CVonline: On-Line Compendium of Computer Vision* 9 (2003).
[8] Meaning Cloud. 2018. Meaning Cloud Topic Extraction. https://www.meaningcloud.com/developer/topics-extraction
[9] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. Emovo corpus: an italian emotional speech database. In *International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA), 3501–3504.
[10] Richard S Croft. 2004. Communication theory. *Eastern Oregon University, La Grande, OR* (2004).
[11] Facebook. 2018. wit.ai. https://wit.ai
[12] Google. 2018. Dialogflow. https://dialogflow.com
[13] Google. 2018. Google Cloud Speech-to-Text. https://cloud.google.com/speech-to-text
[14] Google. 2018. Google Cloud Text-To-Speech. https://cloud.google.com/text-to-speech
[15] Google. 2018. Googlecloud - natural language. https://cloud.google.com/natural-language
[16] Shang Guo, Jonathan Lenchner, Jonathan Connell, Mishal Dholakia, and Hidemasa Muta. 2017. Conversational bootstrapping and other tricks of a concierge robot. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 73–81.
[17] IBM. 2018. IBM ToneAnalyzer. https://www.ibm.com/watson/services/tone-analyzer
[18] IBM. 2018. IBM Watson Speech to Text. https://www.ibm.com/watson/services/speech-to-text
[19] IBM. 2018. Watson. https://www.ibm.com/watson
[20] IBM. 2018. Watson Text to Speech. https://www.ibm.com/watson/services/text-to-speech
[21] Indico.io. 2018. Indico.io. https://indico.io
[22] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155* (2016).
[23] Microsoft. 2018. Azure - text analytics. https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics
[24] Microsoft. 2018. Microsoft Azure. https://azure.microsoft.com/en-us/services/bot-service
[25] Microsoft. 2018. Microsoft Azure Speech to Text. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text
[26] Karen O'Shea, Zuhair Bandar, and Keeley Crockett. 2010. A conversational agent framework using semantic analysis. *International Journal of Intelligent Computing Research (IJICR)* 1, 1/2 (2010).
[27] Rosalind Wright Picard et al. 1995. Affective computing. (1995).
[28] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.
[29] Text processing. 2018. Text processing. http://text-processing.com
[30] Qemotion. 2018. Qemotion. https://www.qemotion.com
[31] Repustate. 2018. Repustate. https://www.repustate.com
[32] ResponsiveVoice. 2018. ResponsiveVoice. https://responsivevoice.org
[33] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*, Vol. 16. 3776–3784.

[34] Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26.
[35] TextRazor. 2018. TextRazor. https://www.textrazor.com
[36] TheySay. 2018. TheySay. http://www.theysay.io
[37] Good Vibrations. 2018. Good Vibrations. https://www.good-vibrations.it
[38] Vokaturi. 2018. Vokaturi. https://vokaturi.com
[39] Joseph Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.