A Modular Data-Driven Architecture for Empathetic Conversational Agents

Conference Paper · January 2021

DOI: 10.1109/BigComp51126.2021.00080

3 authors:

Vincenzo Scotti
Politecnico di Milano
6 PUBLICATIONS **1** CITATION
SEE PROFILE

Roberto Tedesco
Politecnico di Milano
57 PUBLICATIONS **296** CITATIONS
SEE PROFILE

Licia Sbattella
Politecnico di Milano
84 PUBLICATIONS **582** CITATIONS
SEE PROFILE

A Modular Data-Driven Architecture for Empathetic Conversational Agents

Vincenzo Scotti
DEIB
Politecnico di Milano
Milan, Italy
vincenzo.scotti@polimi.it

Roberto Tedesco
DEIB
Politecnico di Milano
Milan, Italy
roberto.tedesco@polimi.it

Licia Sbattella
DEIB
Politecnico di Milano
Milan, Italy
licia.sbattella@polimi.it

***Abstract*—Empathy is a fundamental mechanism of human interactions. As such, it should be an integral part of Human-Computer Interaction systems to make them more relatable. With this work, we focused on conversational scenarios where integrating empathy is crucial to perceive the computer as a human. As a result, we derived the high-level architecture of an Empathetic Conversational Agent we are willing to implement. We relied on theories about artificial empathy to derive the function approximating this mechanism, and selected the conversational aspects to control for an empathetic interaction. In particular, a core empathetic controller manages the empathetic responses, predicting, at each turn, the high-level content of the response. The derived architecture integrates empathy in a task-agnostic manner, hence we can employ it in multiple scenarios by changing the objective of the controller.**

***Index Terms*—Empathetic Computing, Conversational Agents, Deep Learning**

I. Introduction

Empathy is the ability of human beings to *relate another's inner state* [1]. What is perceived can be organised in a hierarchy of aspects, from a physical level to more abstract ones. Emotion is one of the crucial aspects related to empathy; for this reason, *affective computing* theory suggested that we should provide computers with the ability to perceive and display emotions. In this way, they appear to be "genuinely intelligent" and interact naturally with humans [2].

Natural Language Processing (NLP) has been focusing for a long time on the topic of Conversational Agents (CAs) and on techniques to make them as human-like as possible. As a result, in the last years, there has been a rising interest towards Empathetic CAs (ECAs).

However, the models proposed up to now mostly focus only on the emotional aspect [3], [4] while empathy is a more complex phenomenon. Moreover, voice is rarely treated while, in our opinion, it is essential to make ECAs more relatable. On this basis, we argued that considering all the aspects of empathy, and including voice as an essential component, would have helped improving ECA capabilities.

With this work, we introduce the architecture of a data-driven generative ECA, for task-driven Human-Computer Interaction (HCI). With our architecture, we extend the classic modules of a CA [5] to include an *empathetic controller*. This controller works on high-level dialogue attributes related to empathy; depending on what is perceived from the history of dialogue turns, it prescribes the empathetic content of its response, to complete a given task. The architecture we propose is completely modular and task-agnostic.

The rest of the paper is organised as follows. In Section II we present recent solutions in terms of CAs and artificial empathy models. In Section III we describe the modules of our architecture. In Section IV we present the components we're going to employ to build our CA, as well as a brief integration plan. Finally, in Section V we sum up our work.

Partially supported by the European Union's Horizon 2020 project "WorkingAge" (grant agreement No. 826232).

II. Related Works

The agent we have proposed relies on two strong backgrounds. On one side, there are the data-driven CAs; on the other one, the models for artificial empathy.

A. Generative Data-Driven CAs

In the last years, generative CAs for open-domain conversations have begun to gain a lot of interest thanks to the advanced text generation capabilities of Deep Learning models [6], in particular generative systems, which provide more flexibility and adaptivity in the prediction of the response. Up to now, generative solutions have rarely been used for task-oriented dialogues since there weren't ways of controlling the CA responses in an effective way without training on a specific data set. Instead, our proposal tries to adopt a generative model in task-oriented dialogues.

Auto-regressive SEQ2SEQ models [7] based on Transformer [8] architectures currently provide the best solutions for text generation. Thanks to language model pre-training these networks start from a very representative initialization that eases the learning of the conversational capabilities. The resulting CAs can be fine-tuned to include particular functionalities like knowledge or emotional grounding and *persona* consistency. However, current CAs mostly focus on text; voice, if present, is treated by connecting a generic Text-toSpeech (TTS) to the CA textual output. Summing up, these systems provide very good text generation but do not consider empathy at all.

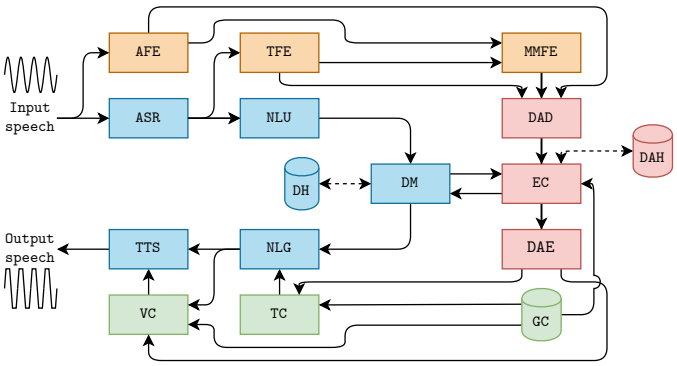

Fig. 1. Modular architecture of the proposed ECA.

Most of the empathetic dialogue systems, on the other hand, focus on rule-based or scripted approaches [9], limiting the expressiveness during the interaction.

With the architecture we have designed, we aimed at dealing with these two problems and cast empathy into task-oriented conversations.

B. Artificial empathy

Recent studies in HCI attempted to formalise empathy for the specific purpose of conversation [1]. The most credited models in this sense are the multi-layered ones. In particular, the Asada's model, appears to be most complete one [10].

This model decomposes empathy along two orthogonal dimensions: the *consciousness* level and the *abstraction* levels.

The abstraction levels are controlled by means of the following dialogic attributes (vocal and/or textual). From the *physical* (i.e., vocal) level, we considered attributes like *speech rate* and *prominence*. At the *intermediate*, emotional, level we have attributes that depend on both voice and semantic content, like *emotion* or *sentiment*. Finally, at the *mental* level we consider attributes that can be seen only through the text, like *dialogue act* or *conversation topic*.

Different kinds of consciousness lead to different empathetic responses, perceivable from the dialogue attributes. Thus, in order to include such dimension, our model identifies *emotional contagion*, *affective and cognitive empathy* and *sympathy and compassion*.

III. Architecture of an ECA

The core architecture we propose, represented in Figure 1, relies on a generative data-driven CA. The classic CA taxonomy locates this kind of agents into the *open domain* CA group [5]. Through the introduction of conditioning modules in the text and speech generation processes, and by means of the EC prescribing the desired output attributes, it is possible to cast this architecture for *task-oriented* conversations. We also introduce empathy by means of a peculiar controller, which works on specific attributes.

A. Core modules

The core CA, represented in blue in Figure 1, is composed of the following modules [5]: the Automatic Speech Recognition (ASR) system converts the spoken input to text; the Natural Language Understanding (NLU), is the component responsible for the understanding of the meaning of the sentence; the Dialogue Manager (DM) controls the structure of the dialogue and maintains the Dialogue History (DH), leveraging the NLU and generating the response (in our case, considering the prescription of the empathetic controller); the Natural Language Generation (NLG) chooses syntactic structures and words to render the meaning of the response; finally, the Text-To-Speech (TTS) synthesis system converts the generated textual response into a speech waveform.

B. Input analysis modules

The input analysis modules, represented in orange in Figure 1, encompass all the components necessary to extract the high-level attributes of an utterance. The Acoustic Features Extractor (AFE) given the raw speech input, computes the features useful to derive the high-level descriptors of the voice pronouncing the utterance. The Textual Features Extractor (TFE) given the raw text input, computes the features useful to derive the high-level descriptors of the utterance transcription. The Multi-Modal Features Extractor (MMFE) module takes care of combining the acoustic and textual features to ease the extraction of descriptors that influence both *linguistic* and *para-linguistic* aspects.

C. Output conditioning modules

The output conditioning modules, represented in green in Figure 1, are the components to be paired with the text and voice generators. They force the response to stick to the controller's prescriptions. The Text Conditioning (TC) module guides the decoding of the response from the NLG given the desired high-level descriptors of the output text. The Voice Conditioning (VC) module guides the synthesis of voice from the TTS given the desired high-level descriptors of the output voice. The Global Context (GC) module holds what we refer to as the *invariant information of the conversation* (e.g. speaker persona and voice timbre). We encode the information within the GC into a continuous embedding representation.

D. Empathetic control modules

The empathetic control modules, represented in red in Figure 1, are the components responsible for modelling empathy. The Dialogue Attributes Decoder (DAD) takes care of projecting the continuous representation given by the extracted features into a discrete representation. The Empathetic Controller (EC) is the core of the ECA; it's the component responsible for selecting the empathetic response to perform a given task. To serve its purpose, the EC decides the output of the current response based on the dialogue inner state as well as the Dialogue Attribute History (DAH) and the GC.

The Dialogue Attributes Encoder (DAE) module provides the opposite function of the DAD. Starting from the discrete attributes predicted by the DAD, the EC projects them into a continuous embedding space. We plan to use these continuous representations to control the response generation.

IV. Development approach

Given the impressive results of Deep Learning models in various NLP tasks, we decided to resort to this framework for all the components of the ECA. In the following, we describe the architectures we selected to implement the ECA modules. Additionally, to conclude, we provide a brief description of the implementation steps.

A. Core modules

The five core CA modules yielded the selection of three separate networks to provide the core dialogue capabilities. We decided to realise the core text-to-text chatbot through a Transformer network. An end-to-end ASR can be attached in input providing the transcription. Similarly, an end-to-end TTS can take the raw text output to pronounce the the response.

NLU, DM (with the DH) and NLG can be all incorporated into a single network, and still provide their functionalities separately. As it has become a common practice nowadays, we decided to fine-tune a Transformer language model since they provide impressive results in terms of text input analysis and output generation.

In particular, we are interested in *Causal Transformer* models, since they've already shown to be compatible with (conditional) dialogue response generation [6]. Additionally, to cope with the common problem of the degradation of long term dialogue information, we have decided to resort to a variant called *Compressive Transformer* [11], which explicitly model a compression function for past tokens to deal with long-range sequences.

For the ASR we considered many choices since there are many networks capable of good accuracy. Among these, we selected Wav2Vec [12], as currently it provides the best results.

For the output voice generation, we have opted for a composition of two networks. The former is the network generating the initial (Mel) spectrogram of the output signal (starting from graphemes or phonemes), we selected Tacotron [13]. The latter, called *vocoder*, is used to refine the raw output signal. For this second network, we have selected WaveNet [14], since it allows for control over speaker voice timbre.

B. Input analysis modules

Input analysis modules are feature extraction networks. For the separate input modalities, we decided to rely on transfer learning from pre-trained networks. These input representations can be further refined by means of additional hidden transformations to build the classification or regression models needed to extract the dialog attributes. Similarly, for the MMFE, additional hidden transformations could merge the separate feature vectors coming from voice and text.

For TFE we adopted *contextual embeddings*, as they reached state-of-the-art performances in many NLP tasks [15]. For this purpose, we can re-use the pre-trained compressive Transformer language model from the core CA.

For AFE we decided to adopt deep convolutional neural networks. In particular, we considered two alternative feature extraction networks that both proved to be useful for voice analysis: VGGish [16] and SoundNet [17].

C. Output conditioning modules

Output conditioning modules permit to control empathetic dialogue attributes. We introduced one module for text control and one for voice control; additionally, there's a module hosting dialogue invariants, encoded in a continuous representation.

For TC we decided to implement a Plug and Play Language Model (PPLM) [18]. The key feature of this model is that it allows to introduce conditioning in text generation without modifying the neural generator and considering also different attributes at the same time.

Differently from the text, speech synthesis has a smaller pool of conditioning solution. The main approach is to pre-train the generator and then separately learn conditioning on embeddings representing the desired attributes [19]. This is the solution we decided to follow for the VC. Additionally, to further refine speech synthesis, we considered augmenting the text to be pronounced with prosodic clues (predicted by the empathetic controller) [20] useful, for example, for putting the focus on specific words.

The GC module, instead, holds the embeddings of invariant information. For text generation, it encodes the representation of the speaker and addressee persona, which are known to be useful for contextual consistency [6]. For voice generation, GC encodes the representation of the speaker's timbre, employed by the TTS to ensure a constant (and selectable) output voice.

D. Empathetic controller

Thanks to the modular architecture, it is possible to design the core EC either in a model-driven or a data-driven way. Since we were interested in data-driven approaches, and particularly Deep Learning ones, we decided to resort to Deep Reinforcement Learning. The controller can be trained through either policy or value algorithms [21] (usually, policy-based algorithms suffer from high variability but working on high-level attributes helps to cope with this problem [22]. To learn the control function, we considered the two main algorithms employed in conversational tasks: *REINFORCE* [23] (for policy learning) and *DQN* [24] (for value learning).

Either way, we need to model the policy function to structure the whole empathetic response according to the model of artificial empathy we are considering; in particular, Equation (1) shows how we represent empathy in the policy function of the reinforcement learning agent. The status s of the *Markov Decision Process* is described by the context c, representing the DAH, and the current inner status of the ECA, given by emotional contagion response π_c and the affective and cognitive empathy response π_e.

We modelled π_c as an *identity function* on the input attributes (the ones compatible with an identity response); in other words, at this level the ECA tries to "stay in sync" with the emotion of the user. We modelled π_e as a *supervised policy function*, learn from large conversational corpora. At this level the ECA tries to manipulate its empathetic response as humans usually do

during generic conversations. At this point, the response is still not aimed for a specific task, it should show understanding of the dialogue content. The sympathy and compassion response π_σ selects the behaviour, depending on the task.

$$\pi_\sigma = \pi_\sigma(s) = \pi_\sigma(\pi_u, \pi_e|c) \quad (1)$$

We composed DAD of a set of classifiers and regressors, designed to extract the discrete empathetic attributes of a dialogue turn. The DAH stores the sequence of attributes of each turn: the sequence used as context by the controller to predict the response. DAE is instead an embedding model, responsible for encoding the sparse and discrete response of the EC into a continuous embedding space. We considered to obtain this last transformation through a function learnt contextually to the conditioning of the output response.

E. Implementation steps

In order to implement the ECA, the idea is to start from a core text-only chatbot, augmenting it with feature extraction and conditioning modules. Then, a first partial version of the controller can be attached. Voice modules can be added on top of this prototype, together with their feature extraction and conditioning modules. Finally, the last part of the controller can be developed. In this way we can estimate the performances of the various components as soon as we integrate them.

For what concerns the data sets for training the various module, we divided them into two groups: *generic* and *labelled*. The former identifies dialogue and speech corpora used to train the core components, they do not require any additional info, only the raw input and output sequences. The latter requires specific labels since they will be used to train both the modules to identify the attributes and condition on them. Thanks to the modular architecture, we do not require a parallel audio-text corpus for training. For the textual part, there exist many data sets of both kinds [25], while to train the voice-related modules labelled data sets are rarer or contain fewer samples. However, some of the information from speech can be extracted automatically, like speech rate or word prominence.

V. Conclusions

In this paper, we have presented a possible architecture for a fully data-driven ECA. We have detailed the modules composing the CA and provided suggestions on the development steps. Finally, we have presented the integration steps we are willing to follow to deploy the ECA. In the future, we are willing to deploy the ECA in conversational scenarios where a more empathetic behaviour is expected from an automatic agent, like automatic psychotherapy sessions. Thanks to the proposed modular architecture, the only piece that would need to be substituted among the different tasks is the core EC.

References

[1] O. N. Yalçin, "Modeling empathy in embodied conversational agents: Extended abstract," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 546–550.

[2] R. W. Picard, *Affective computing*. MIT press, 2000.

[3] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung, "Caire: An empathetic neural chatbot," 2020.

[4] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," 2018.

[5] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009.

[6] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, Apr. 2020.

[7] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[8] Q. Liu, M. J. Kusner, and P. Blunsom, "A survey on contextual embeddings," 2020.

[9] A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth, "Empathy in virtual agents and robots: a survey," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 3, pp. 1–40, 2017.

[10] M. Asada, "Towards artificial empathy," *International Journal of Social Robotics*, vol. 7, no. 1, pp. 19–33, 2015.

[11] J. W. Rae, A. Potapenko, S. M. Jayakumar, and T. P. Lillicrap, "Compressive transformers for long-range sequence modelling," 2019.

[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[13] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[15] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," in *Advances in Neural Information Processing Systems*, 2019, pp. 3266–3280.

[16] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[17] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, 2016, pp. 892–900.

[18] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," 2020.

[19] L. Xue, X. Zhu, X. An, and L. Xie, "A comparison of expressive speech synthesis approaches based on neural network," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data*, ser. ASMMC-MMAC'18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 15–20.

[20] A. Suni, S. Kakouros, M. Vainio, and J. Šimko, "Prosodic prominence and boundaries in sequence-to-sequence speech synthesis," *10th International Conference on Speech Prosody 2020*, May 2020.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[22] C. Sankar and S. Ravi, "Deep reinforcement learning for modeling chit-chat dialog with discrete attributes," in *20th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2019.

[23] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[25] K. Shuster, D. Ju, S. Roller, E. Dinan, Y.-L. Boureau, and J. Weston, "The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents," 2020.