Cyberphysical strategies to develop creative interaction between students and social robots

Manuel Graña, Computational Intelligence Group, Dept. CCIA, University of the Basque Country, Spain

Dialogue systems are becoming central tools in human computer interface systems, moreover in educational environments with social robots. The conventional approaches based on traditional artificial intelligence techniques, have been superseded by machine learning approaches and, more recently, deep learning. In this paper we give a view of the current state of dialogue systems, describing the areas of application, as well as the current technical approaches and challenges. We propose two emerging domains of application of dialog systems that may be highly influential in the near future: storytelling and therapeutic systems.

***Key Words*:** Interactive dialog systems; social robots

1 Introduction

The development of dialogue systems it's been a topic of remarcable interest since the very begining of the Artificial Intelligence [19]. Dialogue systems can be divided into goal-driven systems, such as technical support services, and goal-free systems, such as language learning tools or computer game characters [16]. There has been a long journey since the first conversational system, ELIZA, considered one of the most important chatbot dialog systems in the history of the field [10], to the task-oriented personal assistants that are currently present in most cellphones or home controlers i.e: Siri, Cortana, Alexa, Google Now/Home, etc. This spread of the dialogue systems is linked to the development of a wide range of data-driven machine learning methods have been shown to be effective for natural language processing [15] including the tremendous succed for large vocabulary continuous speech recognition of Deep Neural Networks (DNNs), such as Convolutional Neural Networks (CNNs) and Long-Short Term Memory Recurrent Neural Networks (LSTMs) [14]. Until very recently, most deployed task-oriented dialogue systems used hand-crafted features for the state and action space representations, and require either a large annotated task-specific corpus or a large number of human subjects willing to interact with the unfinished system. This did not only made it expensive and time-consuming to deploy, it also limited its usage to a narrow domain. Conversational systems, however, have drawn inspiration from the use of neural networks in natural language modeling and machine translation tasks [16]. At this point, however, it seems that a link between the two traditionaly separated system development approaches can be achieved, where the task-oriented dialogue systems provide a more natural interaction where there is room for small talk and task-free dialogue. In the same fashion, latest improvements open the door for more complex areas to be approached with this dialogue systems such as therapeutical systems and robot interfaces [6].

Intended contribution The aim of the work in this article is to present the state-of-the-art of dialog systems and some ideas about their future development and new fields of application. The contents of the paper are as follows: Section 2 discusses system architectures. Section 3 comments on the dialog system categories. Section 4 discusses evaluation and training issues. Finally, Section 5 discusses some future challenges.

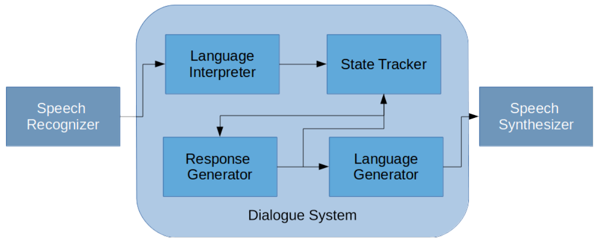

Fig.1 Traditional Dialogue System

2 Architectures of dialogue systems

The traditional architecture for dialogue systems illustrated in Figure 1 includes a series of system modules, each with specific functionality [15]:

- Speech Recognizer, in charge of providing the lexical units for the system extracting them from the voice signal,
- Language Interpreter, in charge of extracting meaning from the stream of lexical units by Natural Language Processing techniques,
- State Tracker, in charge of modeling the dialog state and dynamics, it keeps track of the goal in task oriented systems and of the contextual information in task-free systems.
- Response Generator, produces the semantically grounded response to the current input,
- Language Generator, formulates the response in correct language constructs by Natural Language Generation techniques, and
- Speech Synthesizer generates a recognizable voice signal for the communication with the human side.

Obviusly, the Speech Recognizer and Speech Synthesizer modules have meaning in voice-based dialog systems. Text based dialog systems do not need them. Each of these modules can be tackled with as an independent problem, hence they have been approached using different techniques. This variety is evident in the list of examples in Table 1. Speech Recognition advances due the renew interest in Neural Network are exemplified by [8], or the review in [1]. Deep Neural Networks have been also influential in Language Interpretation [13] and Response Generation [11, 17, 12]. State tracking has been addressed from many sides with a variety of techniques [9].

Fig.2 End-to-end Dialogue System [13]

2.1 End-to-end dialog systems

In recent years the so-called end-to-end dialogue system architectures have become popular and so, some modules, or even all of them, have been collapsed into a unique module of Response Generator, as illustrated in Figure 2. Those systems, mostly based on neural networks, have shown promising results on several dialogue tasks [15]. The main difference between the classical approach and the end-to-end approach is the emphasis on data driven system construction. While the classical approach is much handcrafted and introduces a priori assumptions and design restrictions in all the modules, via specific computational models, the end-to-end approach assumes that the whole architecture can be induced from the data via learning algorithms [4, 3]. This shift has been possible because of the success of Deep Learning Neural Network approaches [7]. There are two main categories of end-to-end dialog approaches [15], on one hand those that searche in a dataset of fixed possible responses, and on the other hand those that select the utterance that maximizes the posterior distribution over all possible utterances. The second approach allows more dynamic responses, as the response generation can be decomposed to the word level.

3 Dialog system categories

There are two basic categories of dialog systems:

- Task oriented systems: these systems have some specific goal that is to be achieved through the dialog interaction. The assistant systems are designed to help search for specific information items. One consequence is that the iteration always reaches a termination state if the user achieves its goals.
- Conversational systems: these task-free systems have not a specific goal, so that iteration can evolve indefinitely, though it is expected that the iterations would produce some evolution of both the user and the cyber-side agent.

3.1 Task oriented dialogue systems

These are the most useful applications of dialog systems: personal assistants where the system needs to understand a request from the user and complete the related task within a limited number of dialogue turns. They are typically designed according to a structured ontology (or a database schema), which defines the domain that the system can talk about[20]. Getting the info is usually achieved using slot-filling, where a dialogue state is a set of slots to be filled during dialoge [3]. However, this system is inherently hard to scale to new domains as it has to be had all features and slots that might be needed manually encoded [3]. Task oriented systems are often designed to carry out some information retrieval dialog [10], such as looking for the nearest restaurant, or to coordinate events, such as planning an appointment or a date.

This kind of systems have beneficed less of the end-to-end architecture and Machine Learning approaches, that do not make assumptions over the domain or dialog state structure [3], because those methods cast the dialogue problem into one of supervised learning, predicting the distribution over possible next utterances given the discourse so far. The supervised learning framework does not account for the intrinsic planning problem that underlies dialogue, i.e. the sequential decision making process, which makes dialogue consistent over time [19].

3.2 Conversational dialogue systems

Conversational aka open dialogue systems try to produce meaningful and coherent responses in the framework of a dialogue history. They have applications ranging from technical support services, to language learning and entertainment, such as playing games with robots [6]. Approaches to build conversational architectures fall into two classes: rule-based systems and corpus-based systems [10]. The rule based systems correspond to the early attemps. such as the famous ELIZA system, where rules were handcrafted following some a priori hints about the desired behavior of the system. On the other hand, corpus-based approaches learn the system structure and parameters from the data in the corpus, making strong use of machine learning and other learning approaches, mining human-to-human conversations, or the human responses extracted from human-machine conversations [10]. Most either rule-based or corpus-based chatbots tend to do very little modeling of the conversational context. Instead they tend to focus on generating a single response turn that is appropriate given the user's immediately previous utterance. For this reason they are often called response generation systems [10]. Given the lack of precise goals, the conversational systems can be formulated as sequence-to-sequence transductors (SEQ2SEQ). However the SEQ2SEQ models tend to generate generic responses, which closes the conversation, or become stuck in an infinite loop of repetitive responses [11].

The most recent computational models used to build the conversational systems are generative models, such as the hierarchical recurrent encoder-decoder (HRED) [16, 17], a kind of Recurrent Neural Networks (RNN) modeling the posterior of the next word in the sequence from the past context by using two contexts, that of the past words and that of the queries performed by the user. The encoder RNN maps each utterance to an utterance vector modeling the hidden state at both contexts, while the decoder RNN models the probability distribution of the utterances conditional to the hidden state. Utterance generation is achieved sampling the posterior probability density.

4 Evaluation and training

System evaluation and training are closely related issues, because the quality measure used for evaluation may be used for training, and the resources employed for evaluation are closely related to the resources employed for training. Some approaches to evaluation use quality measures developed for machine translation systems, such as the bilingual evaluation understudy (BLEU), assuming that the dialog process is akin to a translation process, between the system generated responses and the natural ones from humans. Other use the word perplexity measure [16] from probabilistic word modeling. This approach requires big corpora often unavailable for conversational dialogue systems, and scarce for task oriented systems. Most of the corpora available for dialog system training and tuning come from very specific domains (e.g chats about technical problems such as the Ubuntu IRC chats, or restaurant/movie picking) or were designed for other purposes such as automatic speech recognition system training [15].

Due to the lack of corpora containing precise desired responses for the supervised training of the systems, a natural trend is to resort to Reinforcement Learning (RL) approaches [12, 19, 21], which only require rewards at some point in time, such as the succesful task achievement or some negative rewards when the task-free dialog becomes senseless. The scientific community has turned towards the RL to train and evaluate the dialogue systems since it offers the possibility to treat dialogue design as an optimisation problem, and because RL-based systems can improve their performance over time with experience [5] following a life-long learning approach. However, training dialogue policies in an efficient, scalable and effective way across domains remains an unsolved problem as often requires significant time to explore the state-action space, which is a critical issue when the system is trained on-line with real users where learning costs are expensive[21].

Reinforcemente learning approaches need some mechanism to generate the reward function values. The natural approach is to use human operators that provide rewards according to some quality criteria (i.e. easy of answering, coherence, informativeness, keyword retrieval) but in general it is difficult to extend the approach to wide open dialog systems. A way to automate the process is to apply adversarial approaches [10] mimicking the Turing test of indistinguishability of the machine responses from the human responses. For example in [12] the authors use "a generator (a neural SEQ2SEQ model) that defines the probability of generating a dialogue sequence, and a discriminator analogous to the human evaluator in the Turing test that labels dialogues as human-generated or machine-generated. The generator is driven by the discriminator to generate utterances indistinguishable from human generated dialogs. In the end the human evaluation is the gold standard for all approaches, despite the high cost and inconvenience of having to deal with humans in the loop.

5 Future applications and challenges.

We have tried to illustrate in Figure 3 two emerging domains of application for dialog systems which we have identified as Storytelling and Therapeutics systems. Storytelling is a hybridization of task and conversational systems with many applications in education and entertainment. The interaction is intended to reach the end of the plot, but it can wander along in the path, creating diverging paths that can be creative of new situations. The paradigm of telling a tale while allowing the audience to pose questions and/or ask the audience about their understanding of the current state of the plot and the personages, can be translated also to the teaching of formal concepts and personal training in specific topics in the academic curricula. The dialog system is required to maintain unexpected paths of dialogue and to be able to answer about arbitrarily old states of the dialog or even previous instances of the storytelling process. The system could be adjusted for various degrees of freedom relative to the story and alternative paths leading to the same conclusion of the story.

Therapeutic systems are focused on the user assuming that there is some kind of condition that needs to be reverted or alleviated, which can be pathological in the clinical sense or less dramatical. In the domain of education applications, children showing some aspect of the autistic spectrum can be more accesible to dialog with anthropormophic robots than with humans. In general, the therapeutic dialog system needs to carry out the following tasks, which may or may not correspond to a specific module: diagnostic and evaluation of the user status, selection of treatment, application and assessment of the treatment effects.

Both kinds of innovative dialog systems share the lack or, at best, the scarcity of the available data, because there are no corpuses covering these situations. The model free data drive approaches represented by Machine Learning and Deep Neural Networks may have some difficulties dealing with the need to explain to the medical staff the reasoning leading to some specific treatment and the assessment of the treatment outcomes. The lack of explicit state representation may be an issue when trying to follow divergent paths in storytelling

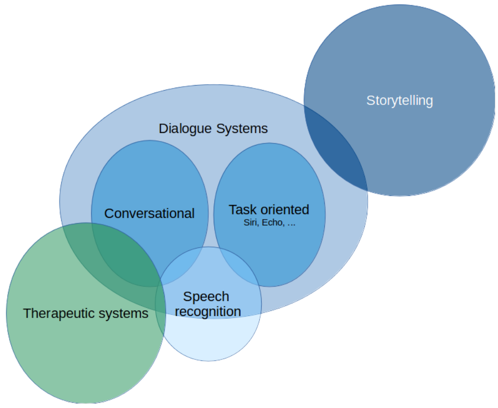

Fig.3 Domains and challenges for dialog systems

or to sharing information with the medical staff. Therefore, new hybridization of the data-driven and the classical dialog architectures may be required.

We will be involved in the development of storytelling systems for educational purposes, specifically the support of children with special needs in the framework of the CybSPEED european project, where we intend to embody these dialog systems in the Nao anthropomorphic robot.

AcknowledgmentsThis project has received funding from the European UnionÕs Horizon 2020 research and innovation programme under the Marie Sk?odowska-Curie grant agreement No 777720

References

[1] Timo Baumann, Casey Kennington, Julian Hough, and David Schlangen. Recognising conversational speech: What an incremental asr should do for a dialogue system and how to get there. In *Dialogues with Social Robots*, pages 421–432. Springer, 2017.

[2] Dario Bertero and Pascale Fung. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135, 2016.

[3] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.

[4] Kevin K Bowden, Shereen Oraby, Amita Misra, Jiaqi Wu, and Stephanie Lukin. Data-driven dialogue systems for social agents. *arXiv preprint arXiv:1709.03190*, 2017.

[5] Heriberto Cuayáhuitl. Simpleds: A simple deep reinforcement learning dialogue system. In *Dialogues with Social Robots*, pages 109–118. Springer, 2017.

[6] Heriberto Cuayahuitl et al. Deep reinforcement learning for conversational robots playing games. 2017.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[8] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[9] Matthew Henderson. Machine learning for dialog state tracking: A review. In *Proc. of The First International Workshop on Machine Learning in Spoken Language Processing*, 2015.

[10] D Jurafsky and JH Martin. Dialog systems and chatbots. *Speech and language processing*, 3, 2014.

[11] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[12] Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.

[13] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):694–707, 2016.

[14] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4580–4584. IEEE, 2015.

[15] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*, 2015.

[16] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784, 2016.

[17] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017.

[18] Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. Reinforcement learning for spoken dialogue systems. In *Advances in Neural Information Processing Systems*, pages 956–962, 2000.

[19] Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*, 2017.

[20] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. On-line active reward learning for policy optimisation in spoken dialogue systems. *arXiv preprint arXiv:1605.07669*, 2016.

[21] Pei-Hao Su, David Vandyke, Milica Gasic, Nikola Mrksic, Tsung-Hsien Wen, and Steve Young. Reward shaping with recurrent neural networks for speeding up on-line policy learning in spoken dialogue systems. *arXiv preprint arXiv:1508.03391*, 2015.

[22] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.

[23] Ji Wu, Miao Li, and Chin-Hui Lee. An entropy minimization framework for goal-driven dialogue management. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.