Conversational Agents, a state-of-the-art review

Luca Molteni
Student of Human Computer Interaction
Academic communication course
Aalto University
Student code: 801775
Email: luca.molteni@aalto.fi

1 Introduction

The purpose of this paper is to present the current state-of-art of Conversational Agents (CA). Conversational Agents are all those dialogue systems capable of carrying out a conversation with a human entity, not only understanding the natural language, but also answering in the same kind of format. In particular we will focus on CA that support speech as an input and reply with synthesized voice.

The rest of the paper is organized as follows. One section will trace the history of CA and how their applications and capabilities have evolved over time. The following sections will be focused on the current state of the technologies used in the various phases of the Natural Language Processing pipeline.

The workflow leading from speech, to understanding, to a vocal reply consists of a number of clearly identifiable phases, as shown in Fig. 1.

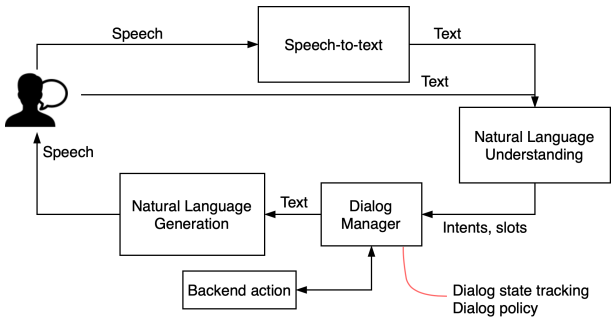

Fig. 1. Information pipeline in task based conversational agents

These phases can be denominated as Speech-To-Text translation, Natural Language Understanding, Dialogue Management and finally Natural Language Generation. For each of these phases a dedicated paragraph will explain the current state-of-the-art.

2 Historical notes

The history of CA traces back to 1966 when Joseph Weizenbaum programmed ELIZA, a chat-bot that answered to textual input mimicking the behavior of a therapist. Most of the time, the program answered simply by reformulating the user's sentences, creating a seemingly impression that it was actually interested to know more. This was due to its extremely simplistic workflow. ELIZA worked by analyzing and substituting simple keywords in a number of predetermined sentences. Sometimes, the result was so effective that people were convinced to talk with a human interlocutor for several minutes.

The first speaking dialogue system was developed during 1977 in the context of the DARPA Project in the USA. [1] The first commercial application was to provide complementary services through phones in the telecommunication industry, e.g., train tables service. Most of subsequent research has focused on developing languages to develop dialogue systems for the industry, such as customer support applications.

In 1995, Richard Wallace released ALICE, a general purpose chatbot, on the same style of ELIZA, but capable of achieving impressive results. One of the biggest improvements compared to its predecessors was the ability to reply with linguistic deflection. These are phrases used to avoid breaking the dialogue when the reply to a certain sentence is unknown, e.g, "Why everything needs to have an explanation?", "Do I really need to answer this question?". Combining this ability with a hugely expanded variety of pre-determined

sentences, ALICE was able to keep a conversation running even when a user tried to ask a few times the same question.

With the advent of Machine Learning techniques and improved computational capabilities, CA started to elaborate speech inputs as well. In 2008, Apple introduced Siri [2], a smart assistant for iPhone devices, able to combine the interactivity of general-purpose dialog systems with the ability to accomplish a variety of tasks, such as searching information on the web, scheduling reminders and initiating calls. From this point in time, the capability of CA to act on the surrounding environment became an essential part of the Human-Computer Interaction.

Siri has been quickly followed by counter-parts from the major competitors, such as Samsung's Bixby and Microsoft's Cortana.

In recent years, smart assistants have moved from our phones to our houses. In 2014, Amazon launched Alexa, a home assistant able to interact not only with the data on the internet but also to several smart devices around the house.

As the precision and capabilities of CA grow, the trend is to entrust them with higher and higher responsibilities such as managing the usage of home appliances or performing payments.

3 Speech to Text

Speech to text models are used to extract text transcriptions from speech input. If for a moment, we move away from the practical implementation, we can usually describe the output of such models as a sequence of graphemes, which are simply the set of symbols of a specific language, such as letter, apostrophes, and punctuation. The input of the model can be speech spectrograms, such as the spectrogram of power normalized audio clips. The models are trained on set of samples with the following characteristics: $(x(1), y(1)), (x(2), y(2)), \ldots$. Each $X(i)$ is a time series of vectors of audio features. For example, in the case of the spectrogram, $X(i)_{t,p}$ can be the normalized power of the p frequency at time t in the audio frame. The output y instead is a textual transcription of the utterance. The goal of the system is at each output time-step t to make a prediction $p(lt|x)$ over characters and symbols denoted by lt. Typically, high-end automatic speech recognition systems are composed of a number of different modules, such as feature extractors or acoustic, language and pronunciation models, needed to take into account the many aspects of human utterances such as noise and accents. The down sight of such systems is that they are very difficult to tune and adapt for different languages or contexts. The latest trend is to build end-to-end systems based on deep-learning models able to implicitly learn a model for the entire automatic speech recognition pipeline. One of the most successful implementation of this philosophy is called Deep Speech 2 [3]. It is a Deep Recurrent Neural Network model made of up to 11 layers, including many bidirectional recurrent layers and convolutional layers. As all models based on DRNN it is very data greedy; this particular model has been trained on around 12000 hours of speech just for English language recognition. Nevertheless, it is able to outperform even human agents on many of the standard datasets.

4 Natural Language Understanding

Natural Language Processing (NLP) is a very broad field that has been developed since the 60s with the work of Noah Chomsky [4]. At a high-level perspective, NLP can be divided in Natural Language Understanding (NLU) and Natural Language Generation (NLG). In this section, we will focus on the former. NLU includes tasks such as Automatic Summarization, Co-Reference Resolution, Discourse Analysis, Machine Translation, Named Entity Recognition, Optical Character Recognition, Part Of Speech Tagging, Sentiment analysis, Intent extraction etc. [5] Out of all these tasks in the context of Conversational Agents we are particularly interested in Intent extraction and Named Entity Recognition (NER). Intent extraction is the process of understanding what the user wants. Imagine that the user asks "What's the weather in London?", the intent here is to get information about the weather in a specific city. In NLU, intent extraction is based on a corpus of example sentences associated to a certain intent. A certain intent can be expressed in different ways, such as "What time is it?" "Can you tell me the time?" etc. Intent extraction algorithms are usually trained on a limited number of intents with a certain number of example sentences for each intent. Given a sentence, the algorithms must be able to predict the most probable associated intent. State of the art models include RNN, Naive Bayes and Support Vector Machines. [6] [7] Named entity recognition instead is the act of assigning phrase nouns and categories to chunked portions of text. For example, from the sentence "Mark drives the yellow bus" it would be identified the chunk "Mark" as a "Person" and "The yellow bus" as a "Vehicle". Also for this kind of task state-of-

art is represented by Convolutional Neural Networks. Currently, a number of commercial and open source services for NLU are available on the market. Between the most famous there are Google's DialogFlow, Microsoft's LUIS, IBM's Watson, Facebook's Wit.ai, Amazon Lex, Recast.ai, Botfuel.io, Snips.ai and Rasa [8].

5 Dialog management

After the NLU process generates a semantic representation of the user input, further processing is carried out by the Dialog Management component. It performs typically two tasks: dialogue state tracking and policy learning. Tracking dialogue states means to estimate the user goal at every turn of the dialogue, combining both the input from the NLU unit and the dialogue history. On the other hand, policy learning deals with the task of choosing the next action based on the dialogue state. Traditional methods include the use of handcrafted rules to select the most likely response [9] or statistical models maintaining a distribution over several possible states of the dialogue such as in [10]. These are the methods generally implemented in commercial products. More recently new research has explored the application of deep learning, reinforcement learning and end-to-end models also for Dialogue Management Components. [11] [12] These new methods have also been driven by the necessity to deal with structural issues of traditional techniques such as the difficulty to adapt to new domains, frequent errors and sensitivity to noise and ambiguity.

6 Natural Language Generation

The last step in the data pipeline of CA is Natural Language Generation. The agent needs to put the chosen action into words by structuring all the necessary information (e.g. retrieved from an external database by the dialogue manager) into a grammatically and semantically correct sentence. NLG needs to generate a specific textual representation from all the possible representations. This step can be split into several subtask such as Content Determination, Text Planning and Sentence Planning. [13] Content determination decides what information include in the output, text planning determines how the information will be structured. Finally, sentence planning deals with deciding how the structure will be divided into sentences and assures that text flows smoothly, e.g., by adding conjunctions, pronominalization and Introducing discourse markers. After the output sentence have been generated it must be transformed into an utterance to complete the interaction cycle with the user. Speech synthesis, which is the artificial production of human speech, can be carried out either by concatenating pieces of recorded speech that are stored in a database or by using a complete vocal model for completely "synthetic" output. The first approach is pretty straightforward, but differences between natural variations in speech and the way in which wave forms have been segmented often result in audible glitches in the output. In recent times new models based on Deep learning-based synthesizers using Deep Neural Networks (DNN) are approaching the quality of the human voice. Examples are WaveNet by DeepMind and Tacotron by Google. [14]

7 Conclusions

In the previous sections of the paper, we briefly ran through the key components and tasks carried out in the information pipeline of Conversational agent. We also reported the most common or effective implementation methods and their recent accomplishments. It is clear that handcrafted rules and custom modules that were the go through solution in past years are being gradually replaced by deep learning models. This is happening mainly because neural networks models are much more flexible and robust and could be also used to implement end to end solution. Task-specific models, on the other hand, are very sensitive to noises and finding the right tuning for all modules require a lot of resources. The greater data consumption of deep models is satisfied by the ever-increasing availability of computational power.

References

[1] KLATT, D., 1976. "Review of arpa speech understanding project". *The Journal of the Acoustical Society of America,* **60**, 11, pp. S10–S10.

[2] Apple machine learning journal.

[3] Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A., Jun, B., LeGresley, P., Lin, L., Narang, S., Ng, A., Ozair, S., Prenger, R., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, Y., Wang, Z., Wang, C., Xiao, B., Yogatama, D., Zhan, J., and Zhu, Z., 2015. Deep speech 2: End-to-end speech recognition in english and mandarin.

[4] Chomsky, N., 1965. *Aspects of the Theory of Syntax*. The M.I.T. Press.

[5] Khurana, D., Koli, A., Khatter, K., and Singh, S., 2017. "Natural language processing: State of the art, current trends and challenges".

[6] Wang, S., and Manning, C., 2012. "Baselines and bigrams: Simple, good sentiment and topic classification". In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, pp. 90–94.

[7] Kim, Y., 2014. "Convolutional neural networks for sentence classification". *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[8] Liu, X., Eshghi, A., Swietojanski, P., and Rieser, V., 2019. Benchmarking natural language understanding services for building conversational agents.

[9] Goddeau, D., Meng, H., Polifroni, J., Seneff, S., and Busayapongchai, S., 1996. "A form-based dialogue manager for spoken language applications". pp. 701 – 704 vol.2.

[10] Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., and Yu, K., 2010. "The hidden information state model: A practical framework for pomdp-based spoken dialogue management". *Computer Speech Language,* **24**, 04, pp. 150–174.

[11] Mrkšić, N., Ó Séaghdha, D., Thomson, B., Gašić, M., Su, P.-H., Vandyke, D., Wen, T.-H., and Young, S., 2015. "Multi-domain dialog state tracking using recurrent neural networks". In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, pp. 794–799.

[12] Wen, T., Gasic, M., Mrksic, N., Rojas-Barahona, L. M., Su, P., Ultes, S., Vandyke, D., and Young, S. J., 2016. "A network-based end-to-end trainable task-oriented dialogue system". *CoRR,* **abs/1604.04562**.

[13] Reiter, E., and Dale, R., 2002. "Building applied natural language generation systems". *Natural Language Engineering,* **3**, 03.

[14] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A., 2017. "Tacotron: A fully end-to-end text-to-speech synthesis model". *CoRR,* **abs/1703.10135**.