Conversational systems: the ultimate goal?

Joseph J Mariani

Conversational systems: the ultimate goal?

Joseph MARIANI
LIMSI-CNRS, Université Paris-Saclay, Orsay, France

Spoken Language Processing has made a huge progress since the first research investigations, 50 years ago, and we now have voice activated systems that entered our everyday life, in cars or at home, such as Google Home, Amazon Echo, MS Homepad or Alibaba Tmall Genie, just to name a few, as they have reach a good enough quality to allow for a large scale deployment.

This has been made possible by scientific research investigations based on machine learning, from pattern matching techniques in the 80s to statistical modeling in the 90s and more recently to neural networks, which got deeper. But this goes also with the availability of large amounts of data to reflect the various languages, dialects, accents, timbers, acoustic environments, applications that can be handled, thanks to the progress in computer storage capacities and processing speed. The introduction of the evaluation paradigm by the US Darpa in the mid 80s has also been decisive in this framework as a major instrument to evaluate the various approaches, select the most efficient ones and compare their technological readiness level with the application needs.

Many of the applications that were foreseen in the early 1980s are now in effect, although this successful black box approach didn't make the human processes of language much better understood. It works, but how and why? The hope is that the available technologies will serve as powerful tools in the hands of researchers in linguistics.

Despite those achievements, there are still scientific issues that have to be solved, and conversational systems still appear as a most challenging goal. I am amazed to see that the topic of the PhD thesis in engineering I defended in 1977, namely pilot-plane dialog, is still unsolved. It appears now as the most difficult challenge to achieve! When looking back, the progresses in research actually aimed at dividing this ultimate goal into sub-problems, which have been solved step by step.

In the PhD thesis in science I defended in 1982, I listed the various levels of speech decoding: the so called “lower” levels, acoustic, phonetic, lexical, and the “upper” ones, syntactic, semantic, pragmatic, dialog, and the same for speech encoding: generation and text-to-speech synthesis. Most of those issues have been successfully addressed nowadays in terms of automatic processing, through acoustic, phonetic and language modeling and semantic slot filling, albeit the pragmatic and dialog ones that are mandatory to achieve conversational systems and are still not available in operational conditions. This includes the handling of anaphora, ellipses, goals, focus, believes, sentiments, metaphors and indirect Speech Acts. I gave in my thesis an example from Barbara Grosz: how to go from a simple affirmation: “The toolbox is locked.” to the “answer”: “The key is in the cupboard”, in a 12 steps reasoning that a normal human is able to conduct easily and instantly, whereas the most powerful machine would still fail. Actually, one may think that the ultimate goal will be completed when a machine will be able to interpret a silence.

The objective and quantitative evaluation paradigm introduced by the US DARPA in the mid 80s has had a decisive impact in allowing progress in our research field, through the organization of yearly evaluation campaigns. It first resulted in moving from knowledge based approaches to pattern matching and machine learning. And it then guided regular progress by identifying sub-problems of gradual difficulty and by checking how well the best system worldwide could solve those sub-problems. It went from read speech (with a limited vocabulary and a reduced language perplexity), to voice dictation (up to an unlimited vocabulary, but facing the practical

difficulty for the user to mentally prepare correct sentences), to the use of varied mikes, to broadcast news transcription, then to conversational speech transcription (Switchboard), including telephone speech and addressing languages other than English, and finally meeting transcription. Text-to-Speech synthesis made progress in parallel that was also measured with both quantitative and qualitative methods, as well as speech understanding for specific tasks, such as flight of restaurant reservations.

Available conversational systems nowadays necessitate a prompt sentence followed by a single request, and even allow to conduct a limited dialog. But they are still unable to conduct a full conversation. While the research benefits from the huge amount of data that is collected everyday by the technology providers which deployed their home devices, one of the existing obstacles is that it is difficult to assess the quality of spoken dialogs systems compared with the evaluation of automatic speech recognition, and that the virtuous loop "trial-test-improve" cannot be implemented as easily as it was for ASR. The reason is that a dialog is dynamic and that there may be an infinite number of correct dialog turns, just like there are a large variety of acceptable translations in Machine Translation that faces the same problem. The relative adequacy of an answer has to be considered together with its fluency, in order to make a dialog both efficient and confortable. There are even attempts for measuring the quality of chatbots and a need for introducing the concept more largely in robotics, in a multimodal communication scheme and a complex environment populated with robots, avatars and networked "things".

Conversational systems may thus appear as the Holy Grail, especially if we consider that conversations should be conducted in any language, and that it therefore implies solving spoken translation or alternatively universal understanding.