Context

*This document is intended to give a general context for a better understanding of the Cezanne-ai project and the research paper: “Cezanne-ai: a conversational AI open-framework for multi-domains, all the languages and limited data”.*

* 1. What is the current context?

1. Classification and probabilistic models worked, but only on simple translation, basic text understanding and short conversations. Deep Learning and Reinforcement learning found some solutions also for more complex texts and longer conversations. One might say that we have the breakthrough and what we need to do is to elaborate on this[[1]](#footnote-1).
2. The scientists and mathematicians came up with these “breakthroughs” and not the linguists, so why not progress with them as main architects, PMs and owners? Of course, specialists in language and English professors were used as consultants and data scientists and this is a good point to go forward, even if it might be a solution to direct the ownership to academicians.
3. The big tech companies have access to unlimited databases which gives them a competitive advantage, but also small companies have seen that no one is interested in regulating learning/training processing and they also can use the pre-trained models provided by big-techs as open-source. This means that data is free for everybody’s use. So, the motto is: let’s process more data (also qualitative) and do more labeling and for sure we will have better results.
4. Some of the DL/RL models were tested also in other domains of AI/NLP with very good results, so why not extend them also to Conversational AI?
5. At DAVOS 2021, Klaus Schwab, the executive president, presented his new book in which he talks about the need for reset in some industry and about the traps provoked by technological progress[[2]](#footnote-2).
6. The complexities of models are increasing every year. From encoder models we went to encoder-decoder, then to bidirectional encoder-decoder and so on… Now we have a Hierarchical Latent Variable Encoder-Decoder[[3]](#footnote-3) and sky’s the limit when you have resources.
7. Microsoft has just patented in the beginning of 2021 a Conversational Bot for dead people[[4]](#footnote-4).
8. We don’t have yet a clear/aligned definition of AGI (Artificial General Intelligence) and there are voices that are saying that maybe after decades we will see positive developments.
9. Inside the top AGI projects only Open AI is covering conversational bots (GPT3) but the models are entirely NLP/DL and are using huge amount of data[[5]](#footnote-5). Other projects are interesting in terms of conversational AI/AGI: COMM AI, BRAIN2BOT, CORE AI, CYC, BAIDU RESEARCH, AGI3, MSR AI, NARS, NEW SAPIENCE, NIGEL, Maluuba, Apple secret AGI project, DUAL, GMU Bica, Google Brain, Hanson Robotics, Humanobs, Real AI, Lisa … but the research made by Baum[[6]](#footnote-6) is showing that the interest is shifting to profit objectives and not to academic research, and for that reason most of the projects are not bringing know-how to conversational bot development.
10. One interesting approach is researched by Inria- Flowers project[[7]](#footnote-7). Rather than trying to imitate the intelligence of adult humans, they follow [an idea formulated sixty years ago by Alan Turing](http://www.loebner.net/Prizef/TuringArticle.html), which actually started to be explored at the beginning of the 21st century: try to reconstruct the processes of human cognitive development in robots, rooted in the dynamical interactions between its brain, its body and its environment, and use this constructive approach to advance our understanding of child development. This approach is called developmental robotics and can be a starting point also for conversational bots including curiosity, autonomy and progressive learning in Machine Education.
11. Genevieve Bell stated in an ATSE article[[8]](#footnote-8): “many ideas were lost in the transition from cybernetics to artificial intelligence, such as the idea of dynamic systems that included technology, people and ecology; the idea of feedback; and the idea that the world you were making should be subject to critical inquiry”. This is a direct judgment to the artificial intelligence limitative principle: *“The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it”.*
12. The Double Bind principle is no longer a concern in AI industry that has stabilized on a steady path, without taking into consideration a need for change/refresh[[9]](#footnote-9).
13. The interest has shifted from research to production, and everybody is waiting to see if the users will be as impressed by the content as they are by the packaging.
14. With the release of BERT paper and open-source code, the path is clear for downstream NLP tasks that can have access to pre-trained models. On Hugging Face we already have thousands of models to be used in different domains, for many tasks and languages. Conversational AI open-domain framework is keeping pace and every day we are seeing that things are moving rapidly. But is it moving in the right direction?

Understanding the context also means responsibilities, but Deep Learning & Co has taken onboard almost everybody, and even most of the academic society is mesmerized (we had informal discussions with linguist involved in NLP) and agrees with the principle that: it doesn’t matter the language, the culture and the customer knowledge and complexity, you only need an embedded lexicon, an E2E algorithm and most of all, a big database and you can fulfill the dream of talking to a robot.

In the beginning, traditional NLP models had focused on the architecture of the natural dialogue (with understanding models playing an important part), but with the data-driven approaches that are more efficient, everybody switched completely (especially in machine translation) to the new thinking, without taking the good parts from the previous findings. We will argue and demonstrate in this paper that this “relative” efficiency is due to pre-training and due to the labeling of the sentences/ phrase (that is fundamentally right) instead of words mapping, but the lack of a proper architecture[[10]](#footnote-10) in the newer models is making it difficult to have a future dynamic in terms of results. The answer could be in the middle, as Mou – 2016 and Liang- 2016 had researched.

* 1. Other considerations:

A conversational bot must fulfill a real need. It is highly debatable if the paths we are now in this area are justifiable even if they uphold Asimov’s laws. Let’s take some examples:

* Example 1. Building social bots to help individuals that have issues with socializing with real people get more comfortable and then help them integrate better in real life social groups. But what if these bots only engage people in virtual discussions and many don’t find meaningful arguments to go back to the real world afterwards?
* 2. Develop virtual personal assistants to make your life much easier. But what if easier doesn’t mean better? Switching the lights in the old fashion way can be better for your body, or making your own decisions based on your own way of making the right selections can be preferable if you have a unique personality.
* 3. Talking with bots that can duplicate a discussion with a dead loved one. But what if the meaning in life is to concentrate on the loved ones that are still alive?
* 4. Building call center/ client service bots that can deal with repetitive questions or situations. This is for sure helping big companies that have many clients but creates even more competitive disadvantages for small businesses (SMEs/SBBs) because standardized solutions, limited databases and infrastructure administration are not effective or/and accessible.

1. NLP specialization – Deepleaning.ai (first 2 modules), <https://www.deeplearning.ai/program/natural-language-processing-specialization/> [↑](#footnote-ref-1)
2. Great Reset – Davos 2021 - https://en.wikipedia.org/wiki/Great\_Reset [↑](#footnote-ref-2)
3. „A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues” by Serban, Sordoni, Lowe..., 2016 [↑](#footnote-ref-3)
4. https://edition.cnn.com/2021/01/27/tech/microsoft-chat-bot-patent/index.html [↑](#footnote-ref-4)
5. Open AI presentation, https://openai.com [↑](#footnote-ref-5)
6. „Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy” by Baum/Fitzgerald/Boddy – 2020 [↑](#footnote-ref-6)
7. Inria- Flowers project presentation, https://flowers.inria.fr/ [↑](#footnote-ref-7)
8. Genevieve Bell, Why we need a cybernetic future: https://www.atse.org.au/news-and-events/article/why-we-need-a-cybernetic-future/ [↑](#footnote-ref-8)
9. https://en.wikipedia.org/wiki/Double\_bind [↑](#footnote-ref-9)
10. Pipelines is the term usually used in these situations [↑](#footnote-ref-10)