Designing a database for a classical theatrical chatbot

Anna Pappa, Samuel Szoniecky, Rocio Berenguer, Joël Huthwohl, Cécile Quach, Arnaud Laborderie

9.3: Designing a database for a classical theatrical chatbot

Authors:
Anna Pappa and Samuel Szoniecky, University Paris 8, France
Rocio Berenguer, Pulso
Joël Huthwohl, Cécile Quach and Arnaud Laborderie, The National Library of France, France

Abstract

The LITTE_BOT project aims to create a theatrical chatbot embodying Dom Juan for Molière's 400th anniversary, presented for the exhibition „Molière, le jeu du vrai et du faux" that the BnF and the Comédie Française devoted to him in 2022. Initiator of the project, Rocio Berenguer, a playwright, approached the French National Library (BnF) to recover the corpus for a literary chatbot. The Gallica Studio project (now closed) encouraged the reuse of Gallica's content, while experimenting with emerging technologies. In this case, exploring voice mediation through chatbots and explaining it to the public. Researchers Anna Pappa and Samuel Szoniecky (both at University Paris 8) contributed their scientific expertise to make the chatbot a reality, in the framework of a call for projects from the Artec university research school.

We aimed to create an open chatbot embodying Molière's Dom Juan. The challenge was to create a database large enough to train the sequence-to-sequence language model. We had to build a database from scratch that would allow an artificial intelligence to imitate Molière's Dom Juan, to speak 17th century French, and to understand the present-day French spoken by its interlocutor.

The challenge for the BnF's Department of Performing Arts was to faithfully recreate Molière for the exhibition audience. Georges Forestier (Sorbonne University) advised the choice of Dom Juan as the ideal character for a chatbot. However, this combination of form and content had a consequence for the database: the play Dom Juan by Molière did not represent a large enough mass. It was therefore necessary to add the Dom Juans of two of Molière's contemporaries, Villiers and Dorimond, and the classical theatre.

The database for training the chatbot is not only the indispensable foundation, but also the most important part of this four-year project. Samuel Szoniecky worked on the semantic analysis of the Molière corpus already encoded in TEI as part of the OBVIL Molière project. The structure of the corpus was analysed (plays, acts, scenes, lines, sentences, keywords) to create items in

an Omeka S database corresponding to each of these structures and their relationships. Rocio Berenguer manually indexed the lines. LITTE_BOT combines two chatbots: an open chatbot based on the Seq2Seq model and a closed chatbot, with indexed lines. However, a few months before the launch, the database was still insufficient, and B12 worked pro bono on the public version of the chatbot using a pre-trained model, GPT-2, and our database.

The empowerment of the public guided certain methodological choices. The project encourages the reuse of Gallica's royalty-free content. The tools and results of the project are returned to the public domain whenever possible. The project is documented so that the exhibition audience can understand how the chatbot works.

The research chatbot continues to be experimented by Paris 8 and Paris Dauphine, with the versions Molière in prose, Molière in rhyme and Brecht. The chatbot created by B12 and hosted in the BOT°PHONE will be reincarnated for other events as other literary characters.

Presenter: *Anna Pappa*

Bio

Anna Pappa received her PHD in Computer Science and since 2006 has been an associate professor of computer science at University Paris 8 in France. Her research focuses on a set of questions related to artificial intelligence and computational linguistics: text analysers, corpus and dataset creation, evaluation and experimentation procedures on corpora, deep learning models for themes such as opinion analysis, automatic annotation (aspects), the construction of specialised lexicons and conversational agents. Her contributions to the field through her research cover different aspects of textual data: collection, analysis, understanding and dialog generation. Anna Pappa has also been involved in collaborative projects with other researchers and industry partners. One of them is the LITTE_BOT collaborative project around conversational agents, chatbots, interacting with characters from 17th century theater. It is a dramaturgy for interacting with the character of

Don Juan, as a part of the perspective of the 400th anniversary of Molière's birth in 2022. My work is about the construction and use of machine learning's dataset and methods, generating dialogues between the human and the machine. The bot gives the cue as if it embodied a character from Molière's theater. Some of the contributions of this work can been found in : "Generative Art Conference – GA2022", "Affects, Compagnons Artificiels et Interactions-ACAI, AFIA 2022", "TALN-RECITAL-ATALA 2022", "Futurs Fantastiques 2021".