MoliAIre, a theatrical agent which speaks like Molière's characters

Guillaume Grosjean, Tristan Cazenave, Baptiste Rozière

LAMSADE, Université Paris Dauphine – PSL, Paris, France

Anna Pappa

LIASD, Université Paris 8, Saint-Denis, France.

Abstract

This project introduces an artistic dimension to automatic conversational agents through the implementation of MoliAIre, a theatrical agent based on a generative language model which gives replicas like Molière's characters. A first version, used as a baseline, shows the artistic potential of the model. We describe the results by exploring two directions that significantly improve the performances of the model. On the first one we train the model on several authors having a similar style to Molière to improve the performances. On the second one, a reverse generation method favours the generation of rhymes. Every model presented in this paper is available online for testing:

https://www.lamsade.dauphine.fr/molierelebot

1. Introduction

Art, like many other fields, is not escaping the growing wave of Artificial Intelligence (AI). The automatic generation of images, texts or music are challenges that guide the research of the AI scientific community and push researchers to develop innovative algorithms. More and more these algorithms become accessible and are now used in many artistic projects.

This project aims to be part of this dynamic by introducing a Conversational Agent (CA) which 'speaks' in the style of Molière's characters, freely accessible by everyone. Beyond the entertainment aspect, this project is the opportunity to introduce CAs in the field of theatre, where we observe a real scarcity on theatrical and literature applications.

2. State of the art

CAs are dialogue systems that can hold an open-domain conversation without specific structure.

Since the introduction of the Transformer architecture [1] and the success of generative language models like *GPT-3* [2], most state-of-the-art CAs are based on an end-to-end decoder-only Transformer neural network. They consist of several billions of parameters and are trained to recreate conversations collected on the web. With *Meena* [3], Google Research showed that a large end-to-end neural model can display more human-like attributes while being simpler than hand-crafted frameworks. FAIR went a step further with *BlenderBot* [4] by

showing that large improvements can be made by fine-tuning on data that emphasizes desirable conversational skills (personality, knowledge, empathy). One of the latest state-of-the-art open-domain CA is *LaMDA* [5], a Transformer-based neural language model with 137B parameters and access to external APIs (information retrieval system, translator, calculator).

3. Baseline model, *MoliAIre*

Following the state of the art, *MoliAIre* is based on a generative language model, pre-trained on a large French text corpus [6], providing it with basic knowledge on French grammar and word meanings. Its weights are then fine-tuned on a corpus of texts, built from Molière's works, each element of which is a dialogue extract of the form:

USER: *line 1*
MoliAIre: *line 2*
USER: *line 3*
MoliAIre: *line 4*
…

Once trained, the model can be used as a CA as follows:

- the user enters a cue. a context identical to the one used for training is created but leaving *MoliAIre*'s cue empty.
- the context is sent as input to the model which completes the cue of *MoliAIre* and stops when a special token representing the end of cue is sampled.
- the process is repeated to generate a conversation by accumulating the exchanges between the user and *MoliAIre.*

Raw data The raw data consists of 32 plays written by Molière, for a total of 15 283 cues.

Making dialogue extracts For each line, a dialogue extract between the user and *MoliAIre* is created by adding subsequent lines from the same scene until the scene is completed or the length of the extract exceeds the maximum length allowed by the model. It is important to make sure that each extract created contains at most two different characters to match the case of a discussion between the user and the model. If it contains more than two characters, it is divided into several sub-extracts containing only two characters. Around 20% of the extracts are removed from the training set to build the validation set.

Pre-trained model We use *GPT-fr*, a *GPT-2*-like model trained on a french corpus with 1 billion parameters, 24 decoder layers, 14 attention heads per layer and a 1792-dimensional embedding.

Training The model is loaded and trained using the Hugging Face python library, on an A6000 GPU with 48Go VRAM. The learning objective is to reconstruct the dialogue extracts by minimizing the logarithm of the perplexity. Each dialogue extract is divided into a sequence of tokens $U = \{u_1, \dots, u_n\}$ and the model parameters Θ are optimized by minimizing:

$$\mathcal{L}(U, \Theta) = -\sum_{i=1}^{n} log\ P(u\_i \mid u_1, \dots, u_{i-1}, \Theta)$$

Inference To favour diversity in the model responses and improvisational

effect, we use a stochastic generation strategy. We choose *top-k sampling* method for generation. According to the context prompt, the model generates the probability over every token to be the next token. We sample the next token among the *k* most likely. We use *k = 40*.

Evaluation The model has a perplexity of 14.88 on the validation set. We introduce a method to use the BLEU score as a style evaluation. BLEU score is an automatic metric that outputs a number between 0 and 100 by comparing candidate texts with reference texts. In practice, the concept of reference text is not easy to define in a theatrical improvisation. The evaluation method tested is the following:

- generate *n* candidate cues with no initial context input to the model to simulate a free cue generation of *MoliAIre*.
- for each generated cue, use the set of original cues from Molière's 32 works as reference text.
- compute the BLEU score between candidates and references.

We found this method to be highly dependent on the generation method and the number *n* of candidate cues. We ran 5 iterations using *top-k* with *k = 40* and *n* = 1000. Results are reported table 1.

Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	**Mean**
30.54	37.99	31.95	29.74	22.92	**30.63**

Table 1 : BLEU score computed for MoliAIre with top-k=40 and n=1000 for 5 iterations.

Given the same input prompt, the model can generate a wide range of cues. It demonstrates a good ability to improvise.

The model doesn't overfit to the training data. Most of the cues it generates can't be found in the original Molière texts. When prompted with an original extract of a scene, the model answers with new lines compared to the rest of the original scene. When prompted with modern French, the model does not change its response style and remains in a style close to that of Molière. The main weakness of the model is its coherence. It is not uncommon for the model to contradict or be inconsistent with the cues entered, especially if the subject evoked is far from the topics addressed by Molière.

4. Directions for improvement

4.1 Multi-author model

In the translation field, a trend in results is that multilingual models can perform better in per-language tasks than their equivalent monolingual model [7].

We studied whether training *MoliAIre* with multiple dramatists improves the performance of the model to generate cues in the style of Molière. We add 502 plays written in the 17th century to the 32 written by Molière, for a total of 189 905 lines. The creation of the dialogue extracts is the same as for *MoliAIre*. The new model, *MoliAIre-2*, is trained in two steps. First, it is trained on the 534 plays set, using a lower learning rate, as a second pre-training task to the generative model. Second, it is trained on the 32 Molière plays set in the same way as *MoliAIre*.

MoliAIre-2 reaches a perplexity of 11.52 on the test set, as *MoliAIre* reaches a perplexity of 14.88 on the same test set. We ran the same BLEU score evaluation and results are reported table 2.

Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	**Mean**
31.02	30.74	48.60	35.49	31.95	**35.56**

Table 2 : BLEU score computed for MoliAIre-2 with top-k=40 and n=1000 for 5 iterations.

While the multi-author model achieves a better BLEU score on average, the score fluctuates a lot between each iteration. It is not trivial to conclude that it is better at predicting and generating cues in the style of Molière than the mono-author model. However, the additional training improves the model coherence, because it has seen more examples of conversations, while preserving Molière style.

4.2 Improving rhymes

MoliAIre and MoliAIre-2 are based on a generative language model that generates text in an autoregressive manner. It makes it difficult to generate rhymes, as the rhyming words are found at the end of the lines.

To improve rhyme generation, we explore reverse language modeling [8]. The main idea is to generate every line in a reverse order, thus sampling the rhyming word at the beginning of the generation of the lines removing constraints of the already generated text.

A new model, *MoliAIre-VERSE*, is trained in the same way as the multi-author model (17th century corpus then only Molière). We filtered the data to train the model on cues in verse only. The difference is in the tokenization process: we reverse the

token sequence of every line, but keep the line order for every dialogue extract, as shown in table 3.

Once trained, the model outputs encouraging results. When the input text is an alexandrine line with rhymes, the model has no trouble generating rhymes thanks to the reversed generation. Moreover, the model seems to have more facility in generating alexandrines, although this is not yet automatic. The rhymes are generally satisfactory, but it happens that the model generates rhymes by using the same word twice.

As the model is trained with verse-only cues, when prompting it with prose cues, it shows more difficulty in generating rhymes.

5. Conclusion

We introduced conversational agents in the field of theatre. We explored two directions that could help improve future related works.

The multi-author model seems to improve the generation in a specific desired style, but a more robust evaluation needs to be conducted to draw real conclusions. A human-based evaluation may be more appropriate to evaluate the performance of the model to mimic Molière's writing style.

The reverse language modelling model shows good performance for rhyme generation. It is an effective and simple method to generate rhymes without the assistance of any external information. Future works could focus on the capacity of such a model to handle both verse and prose cues.

Original cue								
Parbleu,	je	ne	vois	pas,	lorsque	je	m’	examine
1	2	3	4	5	6	7	8	9
Où	prendre	aucun	sujet	d’	avoir	l’	Âme	chagrine
10	11	12	13	14	15	16	17	18
Inverted cue								
examine	m’	je	lorsque	pas,	vois	ne	je	Parbleu,
9	8	7	6	5	4	3	2	1
chagrine	Âme	l’	avoir	d’	sujet	aucun	prendre	Où
18	17	16	15	14	13	12	11	10

Table 3 : Verse inversion method used for MoliAIre-RIME training. Tokens of each verse are inverted but the order of the verses is kept.

6.References

[1] VASWANI et al. (2017). « Attention Is All You Need »

[2] BROWN et al. (2020). « Language Models are Few-Shot Learners »

[3] ADIWARDANA et al. (2020). « Towards a Human-like Open-Domain Chatbot »

[4] ROLLER et al. (2020). « Recipes for building an open-domain chatbot »

[5] THOPPILAN et al. (2022). « LaMDA : Language Models for Dialog Applications »

[6] SIMOULIN et al. (2021). « Un modèle Transformer Génératif Pré-entrainé pour le français »

[7] CONNEAU et al. (2019). « Unsupervised Cross-lingual Representation Learning at Scale »

[8] LO et al. (2022). « GPoeT-2 : A GPT-2 Based Poem Generator »