



HAL
open science

Un chatbot qui donne la réplique !

Anna Pappa, Tristan Cazenave, Guillaume Grosjean, Baptiste Roziere

► **To cite this version:**

Anna Pappa, Tristan Cazenave, Guillaume Grosjean, Baptiste Roziere. Un chatbot qui donne la réplique!. PFIA'22 - ACAI, Jun 2022, Saint-Étienne, France. . hal-04536078

HAL Id: hal-04536078

<https://hal.science/hal-04536078>

Submitted on 7 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Un chatbot qui donne la réplique !



Anna Pappa¹, Guillaume Grosjean², Tristan Cazenave², Baptiste Rozière²

¹LIASD, Université Paris 8, Saint-Denis, France ²LAMSADE, Université Paris Dauphine-PSL, Paris, France



Résumé

A l'occasion du quatre-centième anniversaire de la naissance de Molière (1622-1673), nous présentons un agent conversationnel qui donne la réplique comme un personnage de pièces de théâtre de Molière. Nous avons testé deux modèles de l'architecture Transformer : un qui a été entraîné avec un modèle GPT en français et un corpus de pièces de Molière, et un autre entraîné avec un modèle Seq2Seq sur les oeuvres de Molière uniquement. Les deux modèles parfois manquent de cohérence, mais ils respectent la langue et le style des personnages de Molière lorsqu'ils donnent la réplique. Les deux modèles sont accessibles en ligne :

<https://www.lamsade.dauphine.fr/molierelebot>
<https://www.lamsade.dauphine.fr/seq2seq>

Contexte

L'architecture Transformer [1] puis ses dérivées ont ouvert une nouvelle voie dans la conception des chatbots. Meena [2], basé sur un modèle de langue génératif peut présenter des aspects plus humains qu'un système empirique. Le dernier système LaMDA [3] est basé sur l'architecture Transformer avec 137 milliards de paramètres. Les deux versions du chatbot Molière issues de modèles différents de l'architecture Transformer transposent un personnage des oeuvres de Molière présentant des aspects différents comme la cohérence et la rime.

Problématique

Créer un agent conversationnel qui donne la réplique dans une langue et un style propres à un auteur de théâtre classique, comprendre la langue telle qu'elle était écrite au 17^{ème} siècle, tenir un dialogue cohérent sont les défis qui guident nos recherches. Nous utilisons deux modèles génératifs différents : un modèle seq2seq [4], entraîné uniquement sur les dialogues des oeuvres de Molière, et une adaptation du modèle GPT-2, entraîné sur un corpus français et un jeu de données composé de pièces de Molière pour générer le dialogue entre l'humain et l'agent sans l'aide de systèmes linguistiques complexes.

Méthode

Modèle Seq2Seq Le chatbot *Litte.bot* (projet collaboratif <https://eur-artec.fr/projets/litte.bot/>) est basé sur le modèle Seq2Seq qui permet de répondre à des tâches non structurées, comme le dialogue et la génération de réponses à partir de mots. Dans un des niveaux il encode et décode et dans un autre il met à jour un état de conversation. Nous utilisons une méthode d'apprentissage [4] avec une mémoire LSTM multicouches pour mapper la séquence 'input' à un vecteur de taille fixe, et un autre LSTM pour décoder la séquence 'target' du vecteur.

Modèle GPT Le modèle GPT-fr [5] est pré-entraîné sur un corpus de textes en français. Nous utilisons la version à 1.017 milliards de paramètres. A partir des répliques des oeuvres de Molière, nous créons des extraits de dialogue entre l'utilisateur et *MolièreLeBot* (MLB) suivant la forme :

utilisateur : réplique 1
MLB : réplique 2
...
utilisateur : réplique n-1
MLB : réplique n

Pour chaque extrait de dialogue divisé en une suite de tokens $U = \{u_1, \dots, u_n\}$, nous optimisons les paramètres Θ du modèle en minimisant la log-probabilité suivante :

$$\mathcal{L}(U) = \sum_i \log P(u_i | u_{i-1} \dots u_{i-l}; \Theta)$$

avec l la taille de la fenêtre de contexte. Lors de l'inférence, l'utilisateur entre une réplique et nous créons un contexte similaire à celui de l'entraînement en laissant la réplique de *MolièreLeBot* vide. Le modèle génère la réponse qui est ajoutée au contexte et le processus se répète pour générer une conversation.

Corpus d'entraînement Le corpus d'entraînement est composé de 32 pièces de Molière pour un total de 15240 répliques pour le modèle GPT, et le double pour le modèle Seq2Seq (dédoublé avec les 'target' qui deviennent 'input'). Nous retirons aléatoirement quelques scènes pour construire le corpus de validation (représentant environ 20% des répliques).

Inférence Nous utilisons une méthode de *top-k sampling* pour générer les répliques. En fonction du contexte, le modèle génère la probabilité sur chaque token du vocabulaire d'être le token suivant. L'échantillonnage du prochain token est parmi les k plus probables.

Évaluation

L'évaluation d'un chatbot reste un sujet de recherche ouvert et controversé. De nombreuses métriques automatiques existent (perplexité, BLEU/ROUGE, F1) mais aucune ne semble correspondre parfaitement à l'expérience humaine lors de la conversation avec le chatbot. Le modèle atteint une perplexité de 14.88 sur le corpus de validation. Nous avons essayé de comparer le

score BLEU à différentes étapes de l'entraînement pour quantifier le respect du style de Molière. Nous n'avons pas obtenu de résultats clairs car le BLEU dépend fortement de la stratégie de génération, et le nombre de chevauchements de séquences de n mots entre les répliques de test et les répliques générées ne sont pas faciles à déterminer dans une improvisation théâtrale. Des mesures plus fiables nécessitent une évaluation humaine (SSA [2] ou SSI [3]). Les répliques générées correspondent au style de Molière et à l'aide de la stratégie de génération qui implique un échantillonnage, l'agent peut générer des réponses totalement différentes à partir d'une même réplique.

Exemples

utilisateur : De quoi votre âme est-elle émue ?
MLB : C'est, madame, un sujet qui vient de s'offrir à mon cœur.

utilisateur : De quoi parlez-vous ?
MLB : De la mort.

utilisateur : J'en suis désolée...
MLB : Hé bien oui, c'est une étrange destinée, et je voudrais en souffrir une plus prompte, et plus sûre.

utilisateur : Vous trouverez le bonheur un jour.
MLB : Ah ! Ah ! Voilà ce que j'en dois souffrir.

utilisateur : de quoi votre âme est-elle émue ?

Litte.bot : par le motif d'une reconnaissance cette fille vous sert de toute sa puissance

utilisateur : de quoi parlez-vous ?

Litte.Bot : ne voulant point entendre, il ne répondra pas.

utilisateur : j'en suis désolée ...

Litte.Bot : dieux ! qu'est-ce que j'entends.

utilisateur : vous trouverez le bonheur un jour !

Litte.Bot : je m'en rapporte à vous.

Table – Exemples de dialogues générés avec le *MolièreLeBot* et le *Litte.Bot*, avec les mêmes répliques utilisateur.

Futurs travaux

Nous souhaitons conditionner les générations afin de donner plus d'options à l'utilisateur. Par exemple, nous pourrions faire en sorte que le modèle génère des réponses avec un sentiment imposé (séduction, provocation...) ou un style (prose ou vers). Nous avons besoin d'une métrique basée sur une évaluation humaine pour mieux estimer la performance de nos modèles. Cette métrique pourrait impliquer à la fois experts de Molière pour évaluer le style des réponses générées et le grand public pour évaluer la capacité du modèle à générer des scènes intéressantes.

[1] A. VASWANI, N. SHAZEER, N. PARMAR et al., "Attention Is All You Need," *CoRR*, t. abs/1706.03762, 2017. arXiv : 1706.03762.

[2] D. ADIWARDANA, M. LUONG, D. R. SO et al., "Towards a Human-like Open-Domain Chatbot," *CoRR*, t. abs/2001.09977, 2020. arXiv : 2001.09977.

[3] R. THOPPILAN, D. D. FREITAS, J. HALL et al., "LaMDA : Language Models for Dialog Applications," *CoRR*, t. abs/2201.08239, 2022. arXiv : 2201.08239.

[4] I. SUTSKEVER, O. VINYALS et Q. V. LE, "Sequence to Sequence Learning with Neural Networks," , sér. NIPS'14, Montreal, Canada : MIT Press, 2014, p. 3104-3112.

[5] A. SIMOULIN et B. CRABBÉ, "Un modèle Transformer Génératif Pré-entraîné pour le français," in *TALN, ATALA*, 2021, p. 246-255.