DialoGPT fine-tuned for french language

Emilien VIMONT

September 7, 2021

Abstract

After the advent of more and more powerful models concerning Language Modeling applied to conversational agents, the state-of-the-art today provides extremely strong models in this field. However, most of these models are only applicable to the English language. One of the current challenges is to transpose these models to other languages. I therefore focused here on the efficiency of Fine Tuning a French database on an existing model pre-trained in English (precisely here DialoGPT). This study aims at comparing the performance of Fine Tuning from a perplexity point of view but also from a human evaluation point of view according to different database characteristics. The description and relevance of different databases for Fine-Tuning on the French language will also be discussed.

1 Introduction

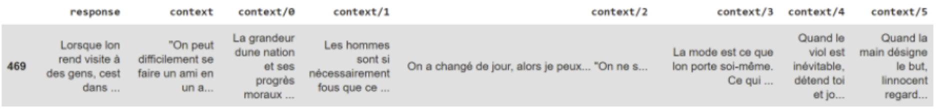
In this work, the first step consists in choosing conversational datasets adapted to the learning of the French language by a pre-trained model in English. Few datasets of this type being available online, the first objective was the creation of clean and usable datasets. Their creation was inspired by the structure of datasets used for pre-training models such as DialoGPT or ParlAI for learning English. The evaluation of the databases is then carried out by calculating the perplexity of each of the models after the training of DialoGPT as well as a human pre-evaluation. The best models are then subjected to a human evaluation.

2 Description of the datasets :

The adaptation performance of DialoGPT to the French language is highly dependent on the database used. In order to best teach the model this language, three conversation databases with different volumes are considered (see Figure 1). In all three cases, the database will be completed with simple dialogues extracted from schoolbook aiming at learning French. (+-500) As few databases of this type are available, most of the databases presented here have been created to best meet the challenges of the task.

2.1 Reddit Scraping

A first approach consists in learning the model on a large volume of data drawn from native speakers talking to each other. The Reddit forum seems particularly suitable for this, as the ease of retrieving this data using the Praw library makes it an ideal choice for a first approach. The data extracted on Reddit and other forums in general are however subject to the social codes used on these forums which can sometimes be incomprehensible to a French speaker who is not familiar with the forum. This



	response	context	context/0	context/1	context/2	context/3	context/4	context/5
469	Lorsque lon rend visite à des gens, cest dans ...	"On peut difficilement se faire un ami en un a...	La grandeur dune nation et ses progrès moraux ...	Les hommes sont si nécessairement fous que ce ...	On a changé de jour, alors je peux... "On ne s...	La mode est ce que lon porte soi-même. Ce qui ...	Quand le viol est inévitable, détend toi et jo...	Quand la main désigne le but, linnocent regard...

Figure 1: Each item in the database therefore consists of a response and 6 previous utterances which act as a context for DialoGPT.

implies a consequent pre-processing work. For this purpose, (1) a list of words to be banned linked to the culture of these forums as well as (2) a list of common words of the English language was made and used as a filter in order to keep only the most acceptable French posts. (3) Posts containing URLs were also removed. It also appeared that the numerous debates related to covid and the health pass in France were "polluting" the database, so some filters were also added to alleviate this problem. The subreddits used for this database are the /france and /rance subreddits which are the most active on the site.

The computational cost of scrapping coupled with the use of numerous filters allows the extraction of a database of up to 15,000 filtered messages. The French Reddit Discussion database (scrapping of 1M discussions on the 5 most active French subreddits dating from 2015) is also studied to perform fine-tuning on a larger amount of data.

2.2 Film and Interview Dataset

A second database is based on the use of subtitles of films and series to which transcripts of culture-related interviews have been added. The subtitles were taken from the OpenSubtitle website and the interview extracts from the CLAPI website, which contains a corpus of transcribed interviews. The advantage of this database is that it has already been subjected to several filters linked to the regulatory bodies of the film industry, thus ensuring a certain cleanliness of the data. However, the format of the subtitles makes it more difficult to extract the words of each character and therefore induces a certain amount of noise in the context of each answer in the database. The presence of punctuation in the interviews and in the subtitles can also be perceived as a witness of data cleanliness but it also induces strange behaviours of the final chatbot when it is confronted with sentences without punctuation (the majority of online conversations).

- Number of conversations : 41719
- Average length of a message : 162.8
- Examples of movies : Guns at Batasi, Je suis un sentimental, I love troubles, The Father, Au revoir là-haut, Lionheart, Human Nature, Flight to Mars, Our Friend, Fontan, Quelque chose à te dire, French Exit, 3 pistole contro Cesare (FR), Faraon, Not as a stranger, Entre les Murs, ...

The following films were chosen in order to preserve a diversity within the corpus, but with a preference for sentimental films in order to maximise conversations rich in emotional exchanges. This may be debatable, however, as a film also relies on images to convey emotions.

2.3 Books dataset

Finally, a last, more qualitative database is composed of extracts from book dialogues. The books chosen are mostly philosophical or spiritual in nature. This database is in opposition to the one based on scrapping, since the dialogues here are quite informative and each answer is quite long. This database is one of the least voluminous and has the lowest perplexity (see figure 2) when used alone. However, the interactions are improved when it is coupled with the film or scrapping database. This is due to the volume of each line of dialogue, interactions tend to be much more chaotic with a model trained on this database alone as the model generates excessively complicated sentences that lose their meaning. The choice of each of the books is made with the knowledge of the work done by OpenAI.

- Number of conversations : 2102
- Average length of a message : 585.1
- Examples of books: Le racisme expliqué à ma fille (Racism explained to my daughter), Le changement climatique expliqué à ma fille (Climate change explained to my daughter), ...

Dataset	Perplexity
Reddit Scraping 2021(12k)	2.29
Reddit Scraping 2021 (15k)	2.40
Reddit Scraping 2021 (15k) Shuffle	2,49
Reddit Scraping 2015 (45k)	5,90
Movies and Interviews Datasets (34k) (*)	1,78
Movies and Interviews Datasets (34k) (Shuffle)	1,82
Books Dataset (1k)	3,06
Books Dataset (2k)	2,31
Reddit 2021 + Books Dataset (18k)	2.26
Reddit 2015 + Books Dataset (62k)	5.18
MBI (Movies, Books and Interviews Dataset) (36k) (*)	1,76
MBIR (Movies, Books, Interviews and Reddit Dataset) (45k)	1,87

Table 1: Mean Perplexity for every model tested (different seeds and learning rate tested (*)). The attribute shuffle means the training sets and test sets have been shuffled before training.

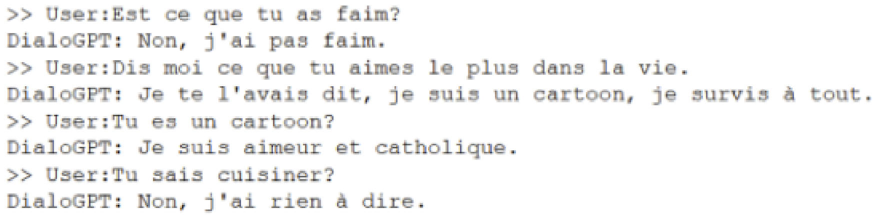
3 Model details

DialoGPT model consists of a large neural conversational reponse generation model. It was pre-trained on 147M conversations extracted from Reddit between 2005 and 2017. DialoGPT provides state-of-the-art results for the English language and gives particularly significant results for the medium size model which will be used here. The model imported is available on HuggingFace. Most of the parameters used for the fine-tuning of DialoGPT were similar to those used by Nathan Cooper in his fine-tuning for the Spanish language. Seed value and learning rate value have however been changed for some models. It has been specified and used to compute the average perplexity score for models in question. The decoder part of the transformer also uses several parameters. After reading previous work and testing the different decoding methods, a top-p and top-K sampling was used for this work with the respecting values of 0,92 and 75. Any further comparisons between these different decoding methods for the specific task of fine-tuning French language has nonetheless not been conducted. It could be a parameter to optimize.

4 Pre-evaluation

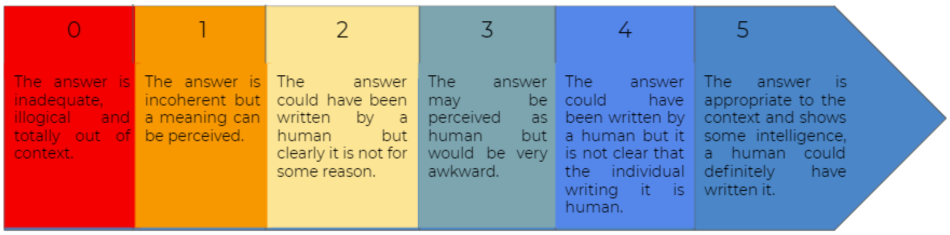
The pre-evaluation of each model is carried out through the evaluation of the perplexity of each model averaged according to different values for the seed and sometimes for the learning rate. This pre-assessment is also carried out by the developer evaluating the model's responses to about ten sentences extracted from a database designed for this purpose. Combinations of several of these databases are also present.

Table 1 shows the results obtained for the different datasets. Firstly, we observe a rather low perplexity for most of the models, this is particularly the case for the book database. This is explained by the density of each of the messages present in this database. However, it is the models with the lowest perplexity that seem to respond best to the evaluation database. However, the results are still quite poor at this pre-validation stage. Most of the responses show an understanding of the function of most of the words in the French language, but it is difficult for the model to formulate intelligible sentences. The book-based model best illustrates this phenomenon. Most of the models thus have a knowledge of French grammar and vocabulary but some models are facing more difficulties than others for this understanding. Perplexity can sometimes be improved by shuffling the context of every message contained in the dataset, this case has been studied in the Table1 with the "Shuffle" attribute. It seems however that this process is inefficient for learning the french language. The MBI dataset has the best score for perplexity but also responds the best to the evaluation dataset, the Reddit dataset from 2021 is also responding well which is why those two datasets are chosen to be submitted to human evaluation.



```
>> User:Est ce que tu as faim?
DialoGPT: Non, j'ai pas faim.
>> User:Dis moi ce que tu aimes le plus dans la vie.
DialoGPT: Je te l'avais dit, je suis un cartoon, je survis à tout.
>> User:Tu es un cartoon?
DialoGPT: Je suis aimeur et catholique.
>> User:Tu sais cuisiner?
DialoGPT: Non, j'ai rien à dire.
```

Figure 2: Extract of a 4-round conversation evaluating the MBI dataset.



0	1	2	3	4	5
The answer is inadequate, illogical and totally out of context.	The answer is incoherent but a meaning can be perceived.	The answer could have been written by a human but clearly it is not for some reason.	The answer may be perceived as human but would be very awkward.	The answer could have been written by a human but it is not clear that the individual writing it is human.	The answer is appropriate to the context and shows some intelligence, a human could definitely have written it.

Figure 3: Humanity rating used by the evaluators of the two models.

5 Human evaluation

The methodology used for evaluating the two models was the following : A first evaluator chatted with the chatbot over a period of 30 utterances. Each utterance was considered independent of the previous one, it was a 1-round conversation. In a second phase, this first evaluator made 32 interactions with the chatbot but this time each interaction was a 4-round conversation. The resulting conversations were then recorded and evaluated by 4 other evaluators. (see Annex and Figure 2) The evaluator had to give a score between 0 and 5 regarding the humanity and the commitment of the answer. (see Figure 3) The grades given for every utterance by the evaluators are then averaged to a final mark for the model.

Results obtained for human evaluation are thus quite low (see Table 2). The evaluators observe that the models tend to understand simple greetings correctly and generate sometimes very accurate answers. However, those accurate answers do not appear as often as they should to consider the models as human. The evaluators also mention that some responses show a lack of comprehension of the message sent by the user, this phenomenon especially appears when the topic of the conversation seems not to be tackled by its dataset. It seems however that a few topics that are not tackled in the datasets are nonetheless answered well by the MBI model. The MBI model is also the only model above the average score of 2,5 but it is not yet sufficient to consider that the fine-tuning gives correct results. The results obtained tend to make one think that larger datasets may give better results for fine-tuning the DialoGPT. The remarks from the evaluators indeed states that the model tends not to understand some situations, a solution could thus be to feed the model with other situations.

6 Reddit scraping evaluation

Finally, Figure 4 shows the different perplexity values obtained for different sizes of filtered Reddit scrapping. We notice firstly that the 2021 scrapping leads to a perplexity more than twice as low as that for a 2015 scrapping, this can be explained by the numerous filters present for the design of the 2021 database. It can be seen that for a database size greater than 50k, the perplexity seems to fall

Model	1-round conversation	4-round conversation
Reddit Scraping 2021 (15k)	2,10	1,89
MBI (Movies, Books and Interviews Dataset) (36k)	2,58	2,37

Table 2: Average score for human evaluation of the MBI dataset and the Reddit scraping from 2021.

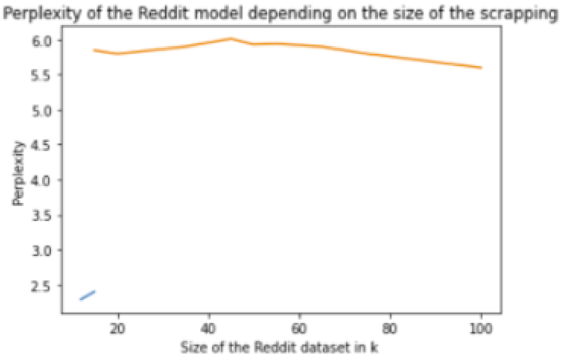


Figure 4: Evolution of the perplexity depending on the size of a Reddit Scraping.

further and further although this value has stabilised at around 5.5. None of the models trained for a database size above 50k can be subjected to human evaluation, however, as these interactions with the evaluation database are very chaotic. Most of the models sometimes respond appropriately to messages but often finish their sentences with 10 times the same phrase. Improvements are observed for the 100k model which repeat the same information for several rounds but introduce it more subtly. A problem with the decoding method was suspected for these specific databases but the changes made (switching to a beam-search, changing the values of top p and top k, temperature) had little impact. Exploiting data from scrapping seems complicated unless perhaps a much larger volume of data is used.

7 Discussion

It seems difficult to obtain state-of-the-art results for the French language using a DialoGPT fine-tuning. It has been shown that the use of different databases, whether dense, small or large (up to 100k for Reddit) only lead to unconvincing results through human evaluation. It appeared during this study that datasets based on forum scrapings are difficult to exploit and often perform less well than more specialised conversations. One avenue to explore for better results may be to further investigate how responses are generated using the decoder, although similar work already exists for the English language. Figure .. shows however that the influence of data volume on perplexity can be significant in the case of a general database such as those obtained using Reddit, increasing volume however implies strange attitudes from the model when tested. Despite the lack of strong results for DialoGPT fine-tuning, it remains that the best results were obtained for the combination of a database based on films, books, interviews and simple dialogues. The fine-tuning of an already existing model for French can therefore be based on a database structure similar to that of this study. Larger datasets built with the same pattern could thus definitely be a way to improve the performances but the relevance of such a process is clearly linked to the amount of data needed for a correct fine-tuning. The objective would therefore be to determine at what volume of data fine-tuning can become interesting. One may then ask whether the development of a pre-trained model in French would not remain more efficient than DIaloGPT fine-tuned in French to obtain state-of-the-art results.

8 Annex

Sociological portrait of evaluators :

Evaluator in charge of discussing with the models :

- Female, 22 years old, Business student.

Evaluators in charge of evaluating the conversations :

- Male, 20 years old, student of mathematics and physics
- Female, 52 years old, Professor.

- Male, 48 years old, researcher in chemistry.
- Male, 22 years old, engineering student.