CLU: Codified Likeness Unit
A conversational bilingual chatbot based on DialoGPT

Mateo Rodriguez
Universita di Trento / Trento, Italy

Abstract

The report discusses the development of CLU, a chatbot designed to emulate an individual's writing style based on their personal WhatsApp data. Initially, the goal was to create an information-explaining chatbot, but due to limitations in scoring and data availability, the focus shifted towards a conversational chatbot that mimics the user's writing style. The project utilized a finetuned version of DialoGPT, a conversational generative model based on the GPT-2 architecture. Training was performed on WhatsApp chats containing mostly English and Spanish conversations. Challenges arose from handling mixed languages within the chats. The results indicated that the base DialoGPT model without Spanish finetuning generated incoherent responses, while the finetuned models only produced partially coherent responses. Future work includes longer training and exploring newer models such as GODEL.

1 Introduction

CLU is a chatbot designed to emulate a specific writing style based on an individual's personal WhatsApp data. CLU is a finetuned version of DialoGPT (Zhang et al., 2020), specifically it uses as the base model an already fine-tuned DialoGPT in Spanish conversations.

The initial idea of the project was to create a chatbot that was able to explain information with the same style of the user, but this idea had two big limitation: scoring and data. The problem was how would it be possible to give a score to a generated output if there's no precise ground data written by the user to which it can be compared. The only data available for training was chit-chat data taken from WhatsApp, thus the idea switched to making a conversational chatbot that mimics the writing style of the user. Before trying DialoGPT, tests were conducted on GPT-2 and GPT-NEO, and both resulted in poor performance, likely due to the language problem explained below, and the extra processing involved in DialoGPT that allows it to work better in a turn based manner.

"Human conversations are also generally more informal, noisy, and, when in the form of textual chat, often contain informal abbreviations or syntactic/lexical errors." (Zhang et al., 2020)

The chat data used contains chats mostly written in Spanish and English, with some few messages in Italian and Portuguese. This is the reason why a Spanish fine-tuned DialoGPT was necessary.

The advantage of the pipeline built is that it can directly process someone else's data thanks to a constant formatting in WhatsApp's exported chats. This means that CLU can be trained to mimic any user that has as a main language Spanish or English, and it can be easily updated/retrained with more recent chats.

2 Related Work

DialoGPT is a conversational generative model created by Microsoft and it's based on GPT-2, which employs a decoder only transformer architecture. It was trained on 147M conversation-like exchanges extracted from Reddit comments, and this allows the model to work in a turn based scenario. This data was filtered, removing any other possible languages by removing comments that didn't include the 50 most frequent English words.

DialoGPT extends the Hugging Face PyTorch transformer, so it can be easily trained and used through the transformers library. The model is available in three sizes:

Model	Parameters	Batch size
Small	117M	128
Medium	345M	64
Large	762M	32

Table 1: DialoGPT model variants

One of the key differences between GTP-2 and

DialoGPT is the use of mutual information maximization which aims to select more informative and diverse responses by penalizing bland or repetitive ones, which is cited as one of the key problems chit-chat chatbots have.

For the Spanish finetuned models there are only the medium (Instituto Tecnológico de Galicia, 2023) and small (Tasar, 2023) variants available and they were finetuned on different data. The small version was finetuned on the OpenSubtitles Dataset, which contains translated movie subtitles. The medium variant was trained on the datasets available in the Bot Framework Tools repository which contains 19314 pairs of messages.

The DialoGPT model as mentioned before uses the GPT-2 architecture as a base, which has already been surpassed in performance by other more updated architectures including it's own successor GODEL.

3 Methodologies

3.1 Dataset

The data from 23 different WhatsApp chats was collected, each with varying lengths. These chats include family members, friends both male and female and also past relationships which represent the main 3 styles of writing. The demanding part of this data is the fact that some chats change languages back and forth at random points, where for example a question was asked in English and the reply is written in Spanish.

The data comes in the following format:

```
08/10/21, 1:10 da tarde - S1: message
08/10/21, 1:10 da tarde - S2: message
08/10/21, 1:10 da tarde - S1: message
08/10/21, 1:10 da tarde - S1: message
message
08/10/21, 1:11 da tarde - S2: message
```

It was necessary to remove all the extra unnecessary formatting that includes the time and data, and who the sender is. Before removing the sender information it was important to join consequent messages as one, separated by a newline character. Some messages, as can be seen in the example above, had no sender or date, these are messages that were sent with new line spacing inside the same bubble. For this it's simply a matter of assigning the writer as the last seen sender.

After doing this we need to format the data into what DialoGPT expects, that is:

message 1 <|*endoftext*|> *message 2* <|*endoftext*|> *message 3* <|*endoftext*|> *message 4* <|*endoftext*|>

There is little documentation on the correct formatting needed for the input data, and some mixed suggestions thanks to the different versions of DialoGPT available, that is the github version, see (Xu, 2020), and the hugginface version. After attempting to use the different configurations the formatting shown above was the one that yielded the best results and was suggested to be used for the hugginface version (Cooper, 2020), (Neskorozhenyi, 2020).

To make the data usable for training it was necessary to group messages into sliding windows of M messages each, and the first message has to always be from the other person, not the user so that the last one is from the user. For example in the case of a window size of 4, the message window N would be composed of messages from 1 to 4, while the message window N+1 would go from messages 3 to 6.

To augment the data the size of the windows was altered randomly, where inside the dataloader a random number is used to reduce the number of messages in a window:

```
split = np.random.randint(M-1)
if split%2 !=0:
    split = 0
lines = self.data[idx].split(EOS)[split:]
```

In the excerpt of code above we take a random number between 0 and 3 (considering M = 4) and if this number is even we split the window and consider the second half of it for training meaning only the last 2 messages.

Using M = 4 and filtering out windows that were longer than 500 tokens the total number of samples available for training ended up being 19168 widows. The data was split as follows:

	Train	Validation	Test
Size	15335	2873	958
Percentage	80%	15%	5%

Table 2: Train, validation and test datasets

Two windows had to be discarded to avoid any information mixing between train and validation/test splits.

Regarding the tokenizer, both models use the GPT-2 tokenizer and it uses the EOS token as the padding token.

Some relevant statistics of the data were also calculated (see appendix A):

Num Tokens	Train	Validation	Test
Mean	101.61	91.51	120.96
Median	81	74	108
Std Dev	71.85	64.34	59.95
90th percnt	499	491	424

Table 3: Dataset statistics

3.2 Training

To train the models two different configurations were used due to hardware limitations. All trainings were done using an RTX 2060 with 12GB of VRAM.

Model	Window	B	Lr	Epochs
Small	4/6	6	$1e{-}7$	25
Medium	4	2	$1e{-}7$	25

Table 4: Train configurations

To speed up training, mixed precision was used. Original trainings were done using floating point 32 and using floating point 16 showed no impact on results. Other relevant aspects of the training: the optimizer of choice was AdamW as it is the one that performs consistently better than others, and the gradient was clipped at 1, to maintain model stability and avoid exploding gradients.

The models were trained on that specific learning rate because using a higher or lower one led to worse performance, this can be seen in the following figure

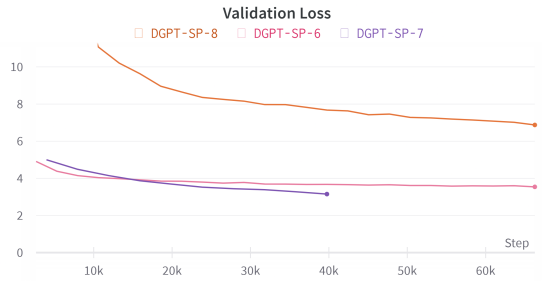

Figure 1: Learning rate comparison

On a side-note, the GPT-2 model (and DialoGPT) from hugginface can take as input the attention mask used to ignore padding, and the token type ids and which are used to mark inputs that belong to different senders. When testing both of these inputs in DialoGPT none of them had any effect on the results.

By considering these factors and configuring the models accordingly, the training process aimed to optimize both performance and efficiency within the given hardware limitations.

4 Results

Previous attempts to get CLU working showed a low training loss but a high validation loss and the generated answers were not coherent whatsoever. This was caused by the formatting of the data, since as said before there are conflicting recommended formats, and because of the model being used was the base DialoGPT model, without any finetuning on Spanish data.

For generating the responses 10 beams were used alongside a temperature of 0.5. The best result from these DialoGPT trains is the following:

Men Jajajajaja io no Ahorita no. Es el cumple de mi hermana y estamos almorzando Qué hora es allá?

Expected response: *dile feliz cumple*
Actual response: *igo estar a estar a de acusar a de acusar a izar algo de acar a izar a izar a izar a izar a izar*

From the example above it's clear that the answer it generated lacks any syntactic coherence, and since the amount of data seemed enough to train a model for such a task the only logical cause for the problem had to be the lack of Spanish training.

For the finetuned models in Spanish, given the limited deterministic methods available to evaluate the model, the scoring was based on the loss, and a likeness score that ranges from 0 to 10 obtained by conducting a survey of 10 replies generated by CLU to 4 different people, from which two of them had their data included in the dataset and two of them didn't. The results were the following:

Model	Train	Valid.	Test	Scoring
Small	5.13	5.04	4.83	2
Medium	1.43	1.34	1.5	3.25

Table 5: Train configurations

The results from both versions of the model we're not satisfactory, the small model didn't present any logical continuation of the conversation, while the medium model barely generated 2 logical correctly spelled answers. The custom scoring shows the same trend as the loss, which

means the loss is correctly evaluating the outputs generated by the model. And it confirms that the medium model performs better than the smaller one.

Some relevant or interesting things that were observed during testing is the fact that the model seems to be stuck repeating the symbols ! and ? and also it produces some partially complete words in Spanish, which is likely caused by the tokenizer being in English.

The best and worst generated conversations from the scoring, alongside the loss graphs can be found in the appendix. All the loss graphs of the trained models and different attempts can be found at https://wandb.ai/mateodrr/CLU.

5 Conclusion

In conclusion, the project aimed to develop CLU, a chatbot capable of emulating an individual's writing style based on their personal WhatsApp data. The initial goal was to create an information-explaining chatbot in the user's style, but limitations in scoring and data led to a shift towards a conversational chatbot that mimics the user's writing style. The project utilized a finetuned version of DialoGPT, a conversational generative model based on the GPT-2 architecture. Training was performed on mostly English and Spanish conversations obtained form WhatsApp chats from various sources. Handling the mixed languages within the chats posed a challenge for the model as it struggles to switch between them.

Two different model configurations were trained, considering hardware limitations. Mixed precision training was used for improved performance. The results showed that the base DialoGPT model without Spanish finetuning generated incoherent responses, and the finetuned models were barely able to write a complete response that was logical and coherent. Scoring was based on loss and a likeness score obtained through a survey.

In summary, the project is the initial step of developing CLU. The medium-sized model showed the best results, but at the end of the day it's likely that the source of the problems is the lack of data or the formatting of it. Other more updated models were tested but they are simply too big to be run locally with the current available resources, one specific model that was of interest is *mosaicml/mpt-7b* since it uses a more updated architecture but it's too big to run.

The source-code can be found at https://github.com/Mateo-drr/CLU

5.1 Future Work

There are two main things that are left for future work. First one is related to the current model, where it would be interesting to train both models for a longer time to achieve a better performance, specially the medium sized model. Also it'd be interesting to apply an extra layer of data augmentation such as *nlpaug.augmenter.char.keyboard* which applies randomly errors in the words by replacing letters with adjacent letters in the keyboard, simulating the common errors in chat data.

The second thing would be to switch to the newer model successor of DialoGPT called GODEL (Peng et al., 2022), which employs an encoder and decoder transformer architecture in order to allow the input of external information (search, documents, databases) to generate a response. This is very interesting because it comes closer to the original idea behind the project, but since it's a newer model there is even less documentation than what's available for DialoGPT which is not much. Apart form there currently no finetuned model in Spanish meaning it'd be necessary to do an initial Spanish finetunning to then start working with the chat data.

References

Nathan Cooper. 2020. Open-dialog chatbots for learning new languages [part 1]. *github and google colab*.

ITG Instituto Tecnológico de Galicia. 2023. Dialogpt-medium-spanish-chitchat. *Hugginface*.

Rostyslav Neskorozhenyi. 2020. Make your own rick sanchez (bot) with transformers and dialogpt fine-tuning. *Towards Data Science*.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. arXiv.

Davut Emre Tasar. 2023. spanish-dialogpt. *Hugginface*.

Xiao Xu. 2020. Fine-tune with own dataset - how to multi-turn ? *github*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

A Appendix

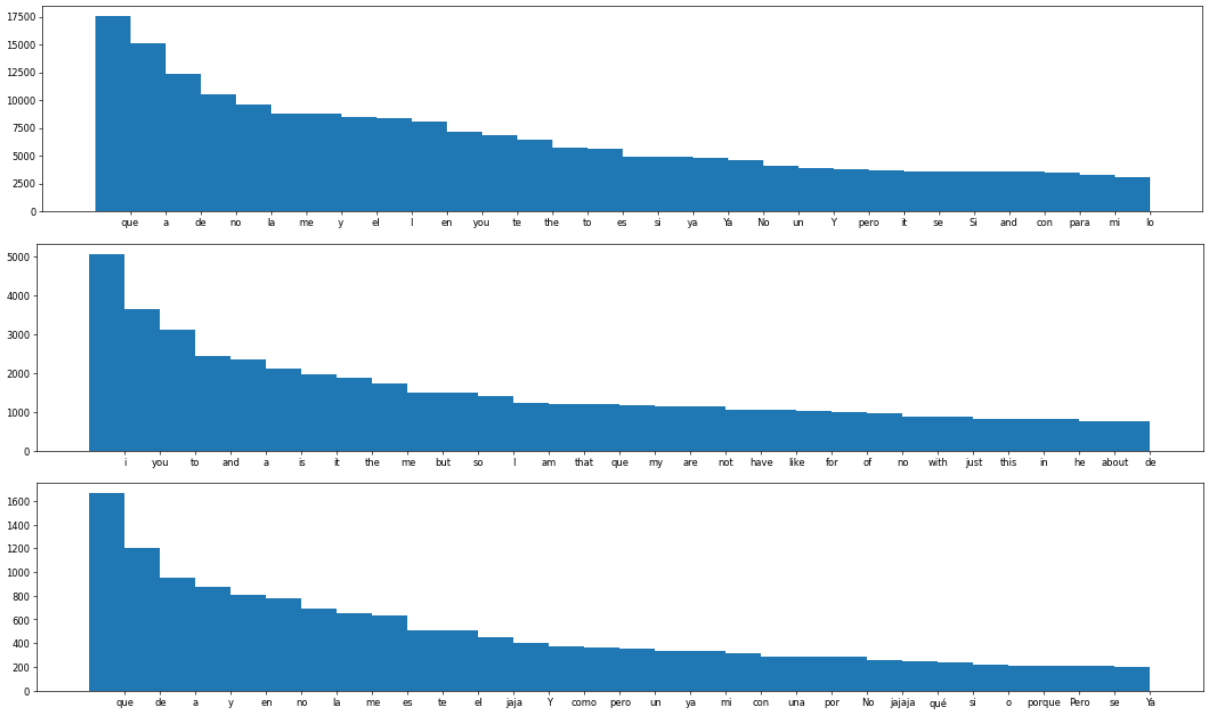

Figure 2: Train, Validation and Test sets word frequencies

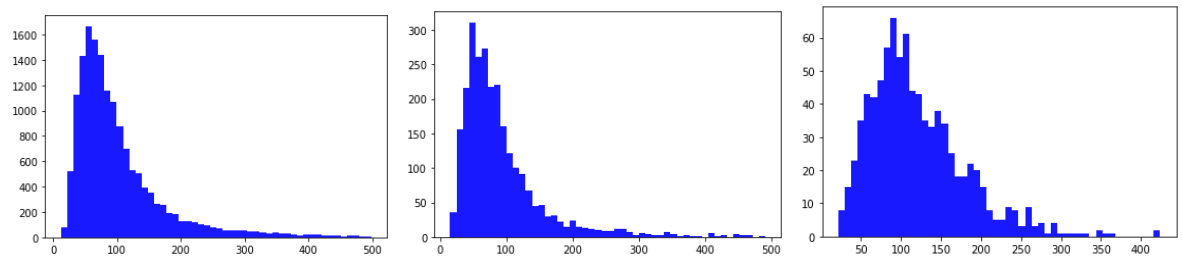

Figure 3: Train, Validation and Test sets tokens length distribution

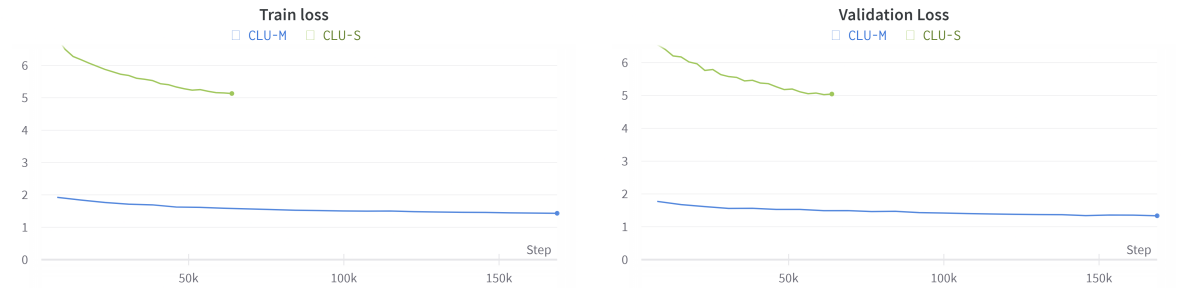

Figure 4: Train and Validation loss

Role	Response
User	So what are you doing tonight
CLU	
User	How are you
CLU	Trying to decide if I should go to sleep or not
User	Oh seems like you are tired
CLU	
User	What is making you feel sleepy
CLU	
User	Do you want to have dinner?
CLU	No
User	Do you want to watch a movie?
CLU	Yes
User	What about star wars?
CLU	How about it
User	Shall we watch star wars?
CLU	Buenas noches?
User	Is Buenas noches a movie?
CLU	gracias.
User	Do you like playing drums?
CLU	irl?

Table 6: Best evaluated chat

Role	Response
User	Que haces. Ya comiste. ?
CLU	
User	Que almorzaste?
CLU	
User	Que vas hacer en la tarde
CLU	?
User	Quieres salí a comer algo
CLU	
User	Vamos a tomar café?
CLU	!
User	O estás muy cansado?
CLU	.
User	Que día es hoy
CLU	lla?
User	Que es eso?
CLU	ilegal
User	Quien es ilegal?
CLU	ial
User	Que es ial?
CLU	ian

Table 7: Worst evaluated chat