Open-domain Conversational Agent based on Pre-trained Transformers for Human-Robot Interaction

Mariana Fidalgo Fernandes[1] and Plinio Moreno[2]
[1]*Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal*
[2]*Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Torre Norte Piso 7, 1049-001 Lisboa, Portugal*

Keywords: Natural Language Processing, Deep Learning, Machine Translation, Transformer, Attention Mechanism.

Abstract: Generative pre-trained transformers belong to the breakthroughs in Natural Language Processing (NLP), allowing Human-Robot Interactions (*e.g.* the creation of an open-domain chatbot). However, a substantial amount of research and available data are in English, causing low-resourced languages to be overlooked. This work addresses this problem for European Portuguese with two options: (i) Translation of the sentences before and after using the model fine-tuned on an English-based dataset, (ii) Translation of the English-based dataset to Portuguese and then fine-tune this model on it. We rely on the DialoGPT (dialogue generative pre-trained transformer), a tunable neural conversational answer generation model that learns the basic skills to conduct a dialogue. We use two sources of evaluation: (i) Metrics for text generation based on uncertainty (*i.e.* perplexity), and similarity between sentences (*i.e.* BLEU, METEOR and ROUGE) and (ii) Human-based evaluation of the sentences. The translation of sentences before and after of the modified DialoGPT model, using the Daily Dialogue dataset led to the best results.

1 INTRODUCTION

NLP is a subfield of linguistics, computer science and AI that employs computational techniques for the purpose of learning, understanding, and producing human language content. NLP recente advances build up on very large amounts of linguistic data and with a much richer understanding of the structure of human language. Efficient and more accurate algorithms that are able to cope the open-domain scenario (Hirschberg and Manning, 2015).

A successful open-domain dialog system requires large amounts of labeled data, which mostly exist for English Language (Ruder, 2020). An approach that leverages on a large English dataset by converting/translating the models to the language in study, provides a viable way to develop a conversational robot with a good performance.

The evolution of Deep Learning (DL) techniques on the automatic translation problem have provided large improvements (Firat et al., 2017). Neural Machine Translation (NMT) relies on deep learning architectures, which considers a broad scope of linguistic sources while looking at whole sentences instead of just words when translating. OPUS MT by (Tiedemann and Thottingal, 2020) belongs to this type of DL-based methods and provides accurate results for the great majority of translated sentences.

The main goal of this work is to develop a chatbot, capable of pursuing conversations on various domains regarding the daily life. The dialogue should be conducted in European Portuguese and should work in a reactive manner. Considering the scarcity of Portuguese dialogues, our approach relies on text translation. Our main building block is an existing model that is fine-tuned in two different manners: (i) A large English database is translated automatically to Portuguese, in order to provide data to a previously trained model in English, and (ii) only the input (from Portuguese to English) and output (from English to Portuguese) utterances were translated while fine-tunning the model.

2 RELATED WORK

The two major categories of conversational AI chatbots are task-oriented and open-domain. On the one hand, task-oriented chatbots have been successfully deployed in several real-life applications, but

Fernandes, M. and Moreno, P.
Open-domain Conversational Agent based on Pre-trained Transformers for Human-Robot Interaction.
DOI: 10.5220/0011300800003277
In *Proceedings of the 3rd International Conference on Deep Learning Theory and Applications (DeLTA 2022)*, pages 168-175
ISBN: 978-989-758-584-5; ISSN: 2184-9277

are limited to their corresponding conversation domain (Dahiya, 2017). On the other hand, open-domain bots aim to serve as a social companion to humans, with whom humans can have engaging and natural conversations. We address the open-domain scenario that needs to master skills such as comprehension, world knowledge, conversation history and constructing valid responses.

In NLP, Words are usually represented as real-valued vectors. For Transformer models, instead of processing words in a sequential manner, all words are processed in parallel, speeding the process and solving the vanishing gradient problem. Transformers use attention mechanisms (Bahdanau et al., 2014) to describe the connections and dependencies of each specific word with all other words in the sentence. The technique that prepares the inputs for a model is also know as tokenizer. The model text input, when encoded, goes through the following pipeline (Wolf et al., 2019): (i) **Normalization:** Operations that involve stripping white space, removing accented characters or lower-casing all text; (ii) **Pre-Tokenization:** It is the act of splitting a text into smaller parts. An intuitive way to think is that this step will divide the text into words; (iii)**The Model:** Once the input is normalized and pre-tokenized, the model has the role of splitting the words into tokens, using the rules it has learned. The tokens will be segments of those words (e.g. "work"+"ing"). It will also map those tokens into their corresponding IDs in the model vocabulary.

The Transformer architecture by (Vaswani et al., 2017), relies on an encoder-decoder model. It leaves recurrence aside and relies on self-attention mechanisms, where the model uses one sequence of symbols enabling it to focus on different words of the sentence and understand its structure.

Language Models are able to predict the next word based on a portion of an utterance. For conversational tasks, the model DialoGPT (Zhang et al., 2019) was pre-trained using multi-turn dialogues extracted from Reddit discussion threads. It is based on the *Open AI*'s GPT-2 (Radford et al., 2019) architecture.

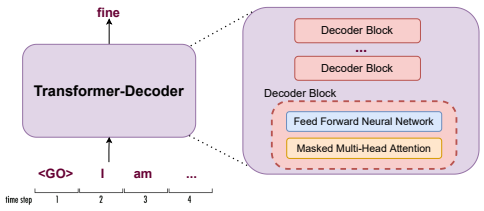

Figure 1: Transformer-Decoder architecture.

The original Transformer (Vaswani et al., 2017) was made up of encoder and decoder blocks also known as transformer blocks. This type of architecture made sense since the model addressed a MT problem, where encoder-decoder architectures (*e.g.* LSTMs, RNN) achieved good results in the past. In later studies, the architecture was stripped of either the encoder or decoder blocks (Liu et al., 2018) to adapt to new tasks, keeping just the Transformer-Decoder blocks, as shown in Figure 1.

The GPT-2 model (Radford et al., 2019) outputs one token at a time and relies just on the mentioned decoder blocks. It is auto-regressive, since that after a token is predicted it is added to the input which will be fed again to the model. In a typical self-attention block, at one position, the model can peak at tokens to its right. However, this is avoided by masked self-attention layers which are used by the GPT-2 model and the DialoGPT. The big setback for this model is that it is only possible to use it pre-trained in English language. Even though it is possible to use it in other languages, the training needs to be done from scratch by the user, which is slow and computationally expensive. Nevertheless, the Transformer architecture is the main block of this work.

NMT achieves great results when a large amount of data is available. However, for low-resource language-pairs, it still remains sub-optimal. This is due to the unavailability of large parallel data, meaning it lacks sentences placed alongside its translations. The advances in ML along with the implementation of NMT techniques impacted significantly the automated translation field (Ranathunga et al., 2021).

OPUS-MT models (Tiedemann and Thottingal, 2020) are trained on state-of-the-art transformer-based NMT. Marian-NMT, which is a stable toolbox with efficient training and decoding capabilities (Junczys-Dowmunt et al., 2018), is applied to the framework. The models are trained on available open source parallel data. Similar to the LMs mentioned before, this system also uses an encoder-decoder architecture with attention mechanisms.

Considering that for every translated sentence, its message is communicated correctly in almost all cases, we can use the OPUS-MT Models translation tool to aid the development of this work.

3 MODELS AND SYSTEM ARCHITECTURE

3.1 System 1 - Fine-tune a Pre-trained Model

In Figure 2, the main steps of the System 1 and the application scenario are shown. The system was pre-

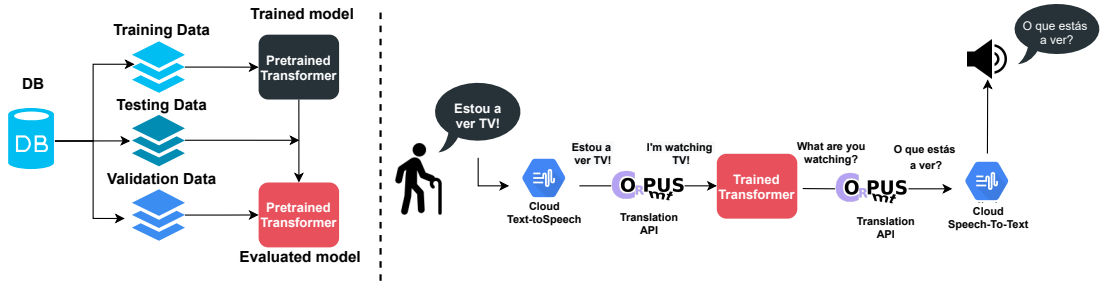

Figure 2: System 1 architecture.

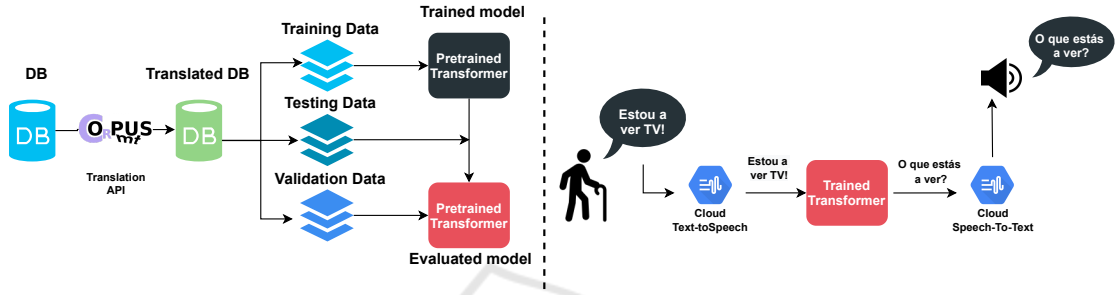

Figure 3: System 2 architecture.

pared with a speech-to-text API, followed by a translation API, which will translate the European Portuguese utterance to English. Afterwards, the translated utterance will be the input for the transformer. The latter will produce the output utterance, that will be translated to European Portuguese and passed through the text-to-speech module. The main idea is to make use of the already pre-trained model which will be fine tuned with the two open-domain conversation databases.

3.2 System 2 - Fine-tune a Pre-trained Model on a New Language

In Figure 3, the second system's architecture is shown. The main idea is to fine-tune the models on a new dataset with a different language. The system is also be prepared with a speech-to-text API and the model is going to be trained with only the small database, due to memory limitations, but translated to European Portuguese. The trained model will receive as input a portuguese sentence and will produce the output portuguese utterance, which will be passed through the text-to-speech module.

3.3 Transfer Learning

Humans have an intrinsic ability to transfer knowledge across tasks. The knowledge acquired while learning one task can be used to solve other related task. The more related, the simpler it is to re-utilize knowledge. This process avoids having to create a network's architecture from scratch and train it during a significant amount of time. Normally, in this technique, a Neural Network is fine-tuned to a specific problem after being trained on a general problem. It allows DL models to converge faster and with less requirements (Malte and Ratadiya, 2019).

Transfer Learning is one of the main foundations for this work. Pre-trained language models, that will be described in the upcoming section, are going to be fine-tuned to address the chatbot problem.

3.4 Dialog Generation Model

GPT-2 works well across a big variety of tasks, mainly due to fine-tuning. This is one of the most used approaches for Transfer Learning, when working with DL models (Guo et al., 2019). It consists in starting with a pre-trained model on the source task and train it further on the desired task. Since the target dataset is small (compared to the one that was used to train the model) and since the number of different parameters is big, if the whole network is fine-tuned to that small dataset, overfitting will occur (Holtzman et al., 2019). The solution to this setback can be to only fine-tune the deep network's final layers, whilst the parameters of the remaining early layers are frozen at their pre-trained values. This solution can be supported due to a combination of both the insufficient training data for the target task and the credible evidence that early layers learn low-level characteris-

tics (Tajbakhsh et al., 2016).

The selected pre-trained model, DialoGPT (Zhang et al., 2019), is fine-tuned with Open AI's GPT-2 (Radford et al., 2019). It is a "tunable neural conversational response generation model" (Zhang et al., 2019) and consists on a GPT-2 model trained on 147M conversation-like data extracted from Reddit comment chains for a 12 year period (2005-2017). Three main models were released with a different number of parameters: small (117M), medium (345M) and large (762M). The model has the purpose to mimic human performance in a single-turn dialogue, acting as a virtual companion for an engaging conversation. The small and medium pre-trained implementations will be tested. Due to limited resources, the smaller model is preferable and it is be able to maintain a coherent dialogue.

3.5 HuggingFace Transformers

The *HuggingFace Transformers* (Wolf et al., 2019) is an open-source library composed of meticulously built state-of-the-art Transformer architectures. It includes tools to ease the training and development of models that have three main parts:

- **A tokenizer** that converts raw text into encodings;
- **A transformer** that takes the encodings and turns them into contextual embeddings, assigning to each word a representation based on its context;
- **A head** that makes predictions, for a specific task, based on the contextual embeddings. (They can be used for fine-tuning or pre-training.)

The objective of text generation is to produce a comprehensible segment of text that follows from the provided context. DialoGPT generates predictions based on the whole conversation history, which is concatenated before it is fed into the model.

4 TRAINING & METRICS

4.1 Model Training

The pre-trained models were tested before any fine-tuning was done, to see the experimental result baseline. These models are very repetitive and could not keep a logical and engaging conversation.

The *HuggingFace* provides the computation of the cross entropy loss. After setting the optimizer (in this case the Adam optimizer), a backward pass is done and an update to the weights is performed.

The fine-tune of the pre-trained model was performed using the two datasets (in English) for System 1 and the smaller one (translated to Portuguese) for System 2. The datasets were converted in a way that every sentence row contains 7 previous utterances for context. It was found that 7 was a good balance between having long enough context to train a conversational model and fit it within the memory constraints (longer contexts take more memory) (Adiwardana and Luong, 2020). The dataset was also tokenized, resorting to the previously mentioned Tokenizers, from *HuggingFace*.

To train the model, a batch of examples is used, with both the inputs and the respective responses. This is due to GPT-2's auto-regressive property, meaning it uses some context to predict the next token. This prediction is then added to the original context and fed back in as the new context for generating the next token.

4.2 Metrics for Evaluation

We consider four common metrics in NLP: Perplexity, question-answer from DialogRPT, BLEU, METEOR and ROUGE scores.

Perplexity: One of the metrics frequently utilized to evaluate the model is perplexity, which measures how unsure the model is in its choice of the following token. The more unsure the model is, the higher its perplexity.

DialogRPT (Gao et al., 2020): Is a set of GPT-2 models trained on 133M pairs of human feedback data (upvotes/replies of dialog systems). The task *human_vs_machine* will be used to attend to this work. The approach was to try to re-rank the fine-tuned DialoGPT outputs with DialogRPT. To calculate these rankings, the Daily Dialogue validation dataset was split by questions and answers. Each part comprises 3870 sentences. The part with the questions is used as the model input. After analysing the DialogRPT methods, a Python script was written that was responsible for:

- Generating 5 different utterances to one input question;
- Outputting the fine-tuned DialoGPT generation probability, for each of those 5 sentences;
- Outputting the DialogRPT ranking probability for each of those 5 sentences;
- Saving one random utterance from the pool of 5, as the answer, for later use;
- Calculating the average generation and ranking probability for all answers.

BLEU-(1, 2, 3, 4), METEOR and ROUGE scores: These metrics use statistical rules to measure the similarity between the output responses and reference re-

sponses. They were initially proposed for MT, however the same idea applies to evaluating generated text as it does to evaluating labels.

The Daily Dialogue validation dataset was split into two. Here, we assume that a normal dialogue has a format of *'question-answer-question-answer...'*, so we have two new sets: Questions and Answers, each with 3870 utterances. Finally, to compute the *Bleu-(1,2,3,4)*, *METEOR* and *ROUGE-L* scores, the Answers dataset was used as reference and the model's predictions to the Questions dataset was used as hypotheses. Even though this may not be the ideally suited method, since more than one reference to the scores computation should be used (paraphrasing), it still enabled a performance contrast among the several experiments. The datasets used for training, testing and validation (questions and answers sets) can be found in the following GitHub repository: https://github.com/marianafidalgo/GrandPal.

Translation Evaluation: We rely on human-based evaluation, which provides a score to the translated dataset. A small interface that collects user evaluation was created. From a dataset that stores 8000 source and target segments pairs, 14 random are displayed to the user at a time. The user then classifies the translation from 1-10.

Chatbot Evaluation: We follow the approach by (Silva, 2020), where users provide answers to statements in a conversation. The conversation consist of utterances (*Training Utterances*) and their corresponding available replies (*Corresponding Robot Utterance*). These replies were provided by 35 people. For the first evaluation method, we tested the different *Training Utterances* and check three scenarios: (i) if the model's answer is equivalent to the *Corresponding Robot Utterance*, (ii) the answer is feasible but does not match correctly and (iii) the answer is wrong.

The second method considers a more realistic human interaction. The user has a chance to engage in a conversation with the trained system. After a six utterance interaction with the system, the user classifies the conversation from 1-10.

5 SYSTEM EVALUATION

5.1 Evaluation of System 1

5.1.1 Daily Dialogue

The first fine-tuned model is the DialoGPT-small on the Daily Dialogue dataset. Some model ablation was done, to analyze the way the configuration variables and hyper parameters affected the model. The memory limitation was inspected, along with the performance of the model, for each different training.

Table 1: DialoGPT-small model with Daily Dialogue dataset.

DialoGPT small - Daily Dialogue													
Nº	BatchSize/ GPU	Grad. Acc.	Epochs	Perplexity	Loss	Gen	Rank	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L
RAW	-	-	-	-	-	57.54%	36.22%	4.00%	1.31%	0.52%	0.23%	4.64%	5.31%
1	2	4	3	11.82	1.36	58.33%	57.31%	4.81%	1.97%	1.07%	0.67%	6.13%	6.73%
2	2	32	3	10.95	1.90	52.89%	54.93%	4.03%	1.54%	0.72%	0.34%	5.71%	6.30%
3	2	32	5	10.99	1.67	55.65%	55.60%	4.52%	1.89%	1.01%	0.60%	6.16%	6.78%
4	4	16	5	9.68	1.44	55.26%	55.83%	4.20%	1.65%	0.83%	0.45%	5.75%	6.25%
5	4	8	5	10.17	1.26	57.57%	56.99%	4.66%	1.83%	0.94%	0.55%	6.08%	6.58%
6	4	12	5	9.81	1.40	56.11%	56.40%	4.59%	1.87%	0.98%	0.55%	5.96%	6.66%
7	2	32	7	10.15	1.45	57.49%	56.96%	4.36%	1.67%	0.85%	0.47%	5.90%	6.25%
8	4	16	7	11.56	1.24	57.06%	56.84%	4.56%	1.81%	0.95%	0.56%	5.92%	6.35%

The first row in Table 1, shows the performance of the raw model with no fine-tuning. We see that fine-tuning a model to a narrower domain leads to an improvement. The red highlighted values point the poorest score, whereas the green highlighted values show the best value, for each metric. In addition, we see that with the increase of the number of epochs, the system tends to overfit. For example in the experiment Nº 8, the loss value decreased, however the perplexity increased significantly.

Finally, the model with the overall best scores is experiment 1. Nonetheless, experiment 4 obtains the lowest perplexity and experiment 3 obtains the best scores for METEOR and ROUGE-L, which provide better insights on the Chatbot task.

The second fine-tuned model is DialoGPT-medium. The memory limitations did not allow to fine-tune models with higher Batch Size/GPU or Gradient Accumulation than the one shown in Table 2.below, failed. The same occurred when the number of epochs was increased. The comparison between the raw model and the fine-tuning experiment can be observed in Figure 2.

Table 2: DialoGPT-medium training results with Daily Dialogue dataset.

DialoGPT medium - Daily Dialogue													
Nº	BatchSize/ GPU	Grad. Acc.	Epochs	Perplexity	Loss	Gen	Rank	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L
RAW	-	-	-	-	-	56.29%	36.88%	4.40%	1.55%	0.67%	0.29%	5.07%	5.46%
1	1	8	3	1.63	0.49	80.29%	60.16%	8.16%	4.93%	3.65%	2.92%	8.52%	10.25%

Figure 4 shows the performance of the model regarding the utterances mentioned in section 4.2.

Note that all answers uttered by the model are written correctly and coherent. The best performance is obtained by DialGPT-small, Experiment 3, even though it is the model with less matched utterances, it is almost always able to answer logically.

5.1.2 Topical Chat

The DialoGPT-small model is fine-tuned, using the biggest English dataset, Topical Chat.The scores of the trained models are shown in Table 3.

GPU limitations did not allow to train the model with more than 2 Batch Size/GPU, or with a greater gradient accumulation than the one shown. The

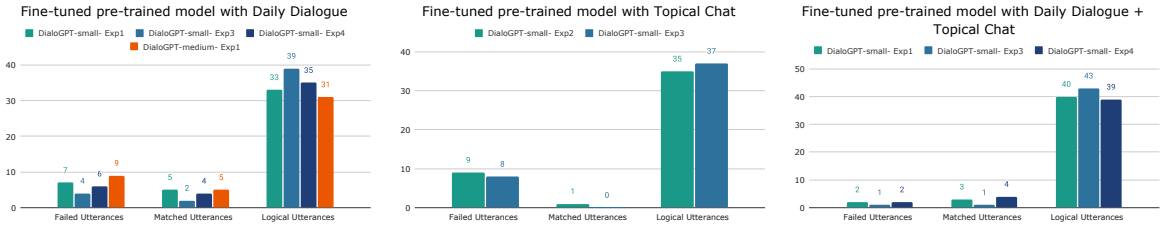

Figure 4: Left side plot shows the results of fine-tuned pre-trained model with Daily Dialogue. Middle plot shows the results of the fine-tuned pre-trained model with Topical Chat. On the right hand side, the results of the fine-tuned pre-trained model with Topical Chat + Daily Dialogue.

Table 3: DialoGPT-small training results with Topical Chat dataset.

DialoGPT small - Topical Chat													
Nº	BatchSize/ GPU	Grad. Acc.	Epochs	Perplexity	Loss	Gen	Rank	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L
RAW	-	-	-	-	-	57.54%	36.22%	4.00%	1.31%	0.52%	0.23%	4.64%	5.31%
1	2	4	3	16.25	2.87	61.72%	64.89%	7.58%	2.18%	0.74%	0.25%	6.32%	6.65%
2	2	16	5	15.00	2.77	61.41%	64.38%	7.59%	2.07%	0.72%	0.27%	6.27%	6.60%
3	2	32	5	14.27	2.66	60.41%	63.50%	7.60%	2.10%	0.73%	0.30%	6.33%	6.66%
4	2	32	7	17.63	1.82	60.41%	65.21%	7.56%	2.04%	0.64%	0.20%	6.30%	6.67%

model with the best result is 3. This experiment shows the most confident model (lowest perplexity), shows good ranking scores, and the best Bleu-1, Bleu-4 and METEOR. Similarly to before, with 7 epochs the model starts to overfit. This can be concluded since it shows the higher perplexity, and also the lowest loss. The evaluation, for the *perplexity* and *loss*, was done using the Topical Chat testing dataset, that contained 11736 utterances. To estimate the rankings and *BLEU*, *METEOR* and *ROUGE-L* scores, the splitted Daily Dialogue validation dataset was used.

Considering Table 3 and Figure 4, we observe that Experiment 3, has the top performance, since it had one less failed predictions and more logical outputs. Due to GPU's memory limitation, the Topical Chat dataset can not be used to train the DialoGPT-medium model.

5.1.3 Daily Dialogue + Topical Chat

We compare the performance of the DialoGPT-small trained on the Daily Dialogue and trained on the Topical Chat. We note that with more data in the training, the model becomes more eloquent and smarter (since it is trained with more information). And even though the model's perplexity increased in the latter, overall the ranking and the metrics scores were better.

Therefore, an experiment was conducted where both datasets, Daily Dialogue and Topical Chat, were merged, resulting in a bigger dataset with a total of 284153 utterances. In order to compute the perplexity and the loss, the model evaluation was done with both testing sets from Daily Dialogue and from Topical Chat. With the results from each dataset, the mean was calculated. The training, with 3 epochs, took around 15h and the one with 5 epochs took 24h. The results are shown below. The attempts to increase the BatchSize/GPU and gradient accumulation failed, again due to the limited GPU. The same occurred when the number of epochs was increased.

Table 4: DialoGPT-small training results with Topical Chat + Daily Dialogue datasets.

DialoGPT small - Daily Dialogue + Topical Chat														
Nº	BatchSize/ GPU	Warmup Steps	Grad. Acc.	Epochs	Perplexity	Loss	Gen	Rank	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L
RAW	-		-	-	-	-	57.54%	36.22%	4.00%	1.31%	0.52%	0.23%	4.64%	5.31%
1	2	200	16	3	13.05	2.56	65.47%	53.21%	3.90%	1.78%	1.00%	0.60%	6.19%	6.99%
2	2	200	32	3	13.23	2.57	64.13%	52.70%	3.83%	1.64%	0.93%	0.56%	5.94%	6.72%
3	2	200	16	5	13.54	2.59	68.11%	55.50%	4.75%	2.39%	1.52%	1.04%	6.88%	8.07%
4	2	200	32	5	13.13	2.59	66.13%	54.10%	4.05%	1.87%	1.10%	0.71%	6.30%	7.14%

Comparing the results from Table 4 and the right side of Figure 4, we see that the best generation probabilities and ROUGE-L scores were obtained with this dataset. Note that with a bigger dataset, the model is able to generalize more, learning how to dialogue in a wider domain. Despite the fact that the scores are not the best in comparison to the previous experiments, the model from experiment Nº 3 was the one that had the best performance when conducting a dialogue, only failing 1 utterance. It was not possible to use this big dataset to fine-tune the DialoGPT-medium model due to the GPU constraint.

5.2 EVALUATION OF SYSTEM 2

5.2.1 Translated Daily Dialogue

We use the OPUS-MT model for translating the Daily Dialogue datasets. All the training with a higher Batch Size/GPU or Gradient Accumulation than the one shown in Table 5, failed. The same occurred when the number of epochs was increased.

The metrics were calculated using the translated Daily Dialogue testing dataset, containing 8069 utterances. For each experiment, the *Bleu-(1,2,3,4)*, *METEOR* and *ROUGE-L* scores were computed.

The first two rows in Table 5, shows the performance of the raw model with no fine-tuning with the english Daily Dialogue validation dataset and with the translated Daily Dialogue validation dataset.

Note that the Bleu-1 score, from the RAW-PT model is higher than the one from the fine-tuned

Table 5: DialoGPT-small training results with translated Daily Dialogue dataset.

DialoGPT small - Translated Daily Dialogue												
Nº	BatchSize/ GPU	Warmup Steps	Grad. Acc.	Epochs	Perplexity	Loss	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L
RAW	-	-	-	-	-	-	4.00%	1.31%	0.52%	0.23%	4.64%	5.31%
RAW-PT	-	-	-	-	-	-	3.37%	0.63%	0.12%	0.00%	2.66%	3.33%
1	2	200	8	3	6.87	1.93	3.02%	1.33%	0.72%	0.41%	3.64%	5.18%
2	2	200	16	3	7.22	2.05	2.36%	0.87%	0.40%	0.21%	3.16%	3.97%

model, due to the repetitiveness of the RAW-PT. For that reason, if the model knows how to utter some portuguese words and those words are in the reference, the score increases as many times as they appear.

To the same input question **"Sim, tenho um entendimento geral."**, the reference answer and the RAW-PT model prediction were compared. **Answer reference:** Eu acho que você já tem bom conhecimento sobre a nossa empresa. **Model Hypothesis:** Eu n o fazia eu acho que acho que eu n o fazia eu acho que eu acho que eu acho que eu acho que acho que acho que eu acho que acho que eu acho que acho que acho que eu acho que.

Analysing the rest of the Table 5, the model that performed best was Nº 1, however the metrics are still poorer when compared to the RAW model. Table 5 and Figure 5 show that this model has the worst performance in comparison to the previous models. By training the model with a small, translated dataset, one layer of error is already being added to the system. This is due to the fact that the translation is not flawless.

The DialoGPT-small model trained on the Daily Dialogue has some issues. By translating the dataset, the performance only decreased. However, we evaluate the approach that consists in translating input and output of the model that performed best: DialoGPT-small model, fine-tuned on the merged dataset (Daily Dialogue + Topical Chat), experiment 3.

5.2.2 Machine Translation

To evaluate the MT model, the human classifications from the interface mentioned before were analyzed. From the 8000 utterances, 784 were evaluated. Taking into consideration that 14 sentences are displayed per page, there were 56 evaluated pages. Counting with, approximately, 2 pages per person, we estimate a participation of 30 people. The results are shonw in the middle plot of Figure 5. This grading goes in accordance to the belief mentioned that for every translated sentence, its message is communicated correctly in almost all cases, since most of the translations obtained the maximum grade.

6 PIPELINE EVALUATION

The full system was evaluated in two realistic conversation scenarios. We choose the best model from the above experiments (Transformer from experiment 3 of System 1, trained on the merged dataset -Daily Dialogue + Topical Chat) and the Translation API was OPUS-MT.

The first experiment follows (Silva, 2020). The portuguese utterances were inputed into the system, with the translation API for the input and output. The prediction was compared with the *Corresponding Robot Utterance*. From the 45 given utterances given to the system, 2 were wrongly translated but only 1 fell out of context. None of the utterances matched the *Corresponding Robot Utterance*, and the remaining sentences were logical answers.

In the second experiment, people evaluated the answers provided by the selected system. We counted with the participation of around 30 people resulting in 40 conducted dialogues. The results are shown on the right side of Figure 5. This grading means that the Chatbot is able to conduct a fair dialogue, being 7 the rating with more occurrence.

7 CONCLUSION

We propose an approach for a conversational Portuguese robot based on the State-of-the-Art Transformer. Our solution generates a robot utterance for every human utterance, allowing a coherent conversation between a person and a robot. Nevertheless, deciding on a DL path brings some challenges such as: (i) The need of having huge amount of data for the model training, (ii) the need of a good graphics card (GPU) with a considerable memory size, (iii) the lack of data specialized in elderly dialogue. We address these challenges by using Transfer Learning, introducing an open-domain chatbot that was trained with limited GPU resources and on a language with few dialogue resources (European Portuguese).

All the trained models obtained better results than their corresponding baselines. The most adequate model for human-robot conversation is the model DialoGPT, fine-tuned in the English-based (formed by Daily Dialogue with Topical Chat) dataset, with a translation layer to the input and output.

The robot responds to the last human phrase uttered. It also considers the course of the conversation, due to the introduction of the last human and robot utterances, as well as the history of human inputs to the model and the robot predictions.

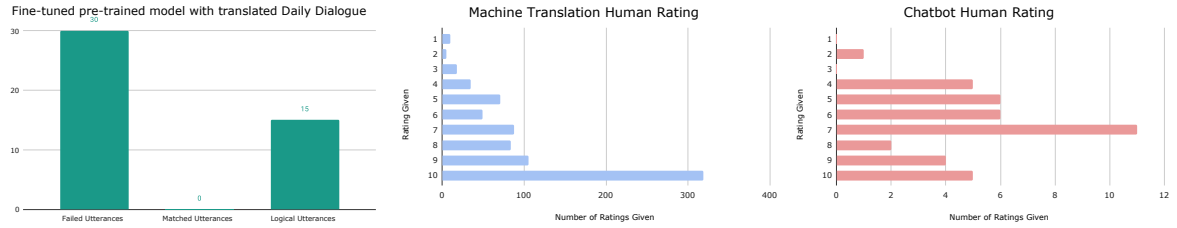

Figure 5: Lef side plot shows the fine-tuned pre-trained model with translated Daily Dialogue. Middle plot shows the Human Evaluation of the Machine Translation. Right side plot shows the Human Evaluation to the System

ACKNOWLEDGEMENTS

This publication has been partially funded by the project LARSyS - FCT Project UIDB/50009/2020 and the project and by the project IntelligentCare – Intelligent Multimorbidity Managment System (Reference LISBOA-01-0247-FEDER-045948), which is co-financed by the ERDF – European Regional Develpment Fund through the Lisbon Portugal Regional Operational Program – LISBOA 2020 and by the Portuguese Foundation for Science and Technology – FCT under CMU Portugal.

REFERENCES

Adiwardana, D. and Luong, T. (2020). Towards a conversational agent that can chat about... anything. *Google AI Blog*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Dahiya, M. (2017). A tool of conversation: Chatbot. *International Journal of Computer Sciences and Engineering*, 5(5):158–161.

Firat, O., Cho, K., Sankaran, B., Vural, F. T. Y., and Bengio, Y. (2017). Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Gao, X., Zhang, Y., Galley, M., Brockett, C., and Dolan, B. (2020). Dialogue response ranking training with large-scale human feedback data. *arXiv preprint arXiv:2009.06978*.

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., and Feris, R. (2019). Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4805–4814.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., et al. (2018). Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Malte, A. and Ratadiya, P. (2019). Evolution of transfer learning in natural language processing. *arXiv preprint arXiv:1910.07370*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., and Kaur, R. (2021). Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Ruder, S. (2020). Why You Should Do NLP Beyond English. http://ruder.io/nlp-beyond-english.

Silva, C. (2020). Natural Language Processing: My "grandchild-Bot".

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

Tiedemann, J. and Thottingal, S. (2020). Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.