Speech Processing: Lab 3 Report

Antonio Vitor Villas Boas[1], *Rita Soares Leite*[2]

MECD, IST[1], MECD, IST[2]

ist1105429, ist192646

1. Introduction

The goal of this laboratory is to develop an end-to-end dialogue system, composed by three tasks. The first task consists of a speech-to-text model, where an Automatic Speech Recognition model (ASR) is used to convert the user's voice signal into text. The second one is a text-to-text, where we generate a response for the previously obtained text using large pre-trained language models (LLM). Finally, the system must convert the generated response into a synthetic speech signal.

In the following sections we will present a thorough analysis of the techniques used and their respective results, for each individual tasks, as well as some examples to illustrate the performance of the complete dialogue system.

2. Speech to Text

As was already explained above the first task consisted of constructing a speech to text pipeline, capable of receiving an audio signal, and returning a transcript of the spoken words. With this in mind, we used two models and Word Error Rate (WER) as metric to evaluate which one performed best. The first model was the HuggingFace Automatic speech recognition [1] and the second was OpenAI Whisper [2], a pre-trained model for automatic speech recognition (ASR) and speech translation.

The group recorded 2 little stories - about 30 seconds long - to be transformed into speech. Each model converted both stories to text. Table1 shows the WER scores of each model.

Pipeline ASR	Whisper
1.01575	0.53543

Table 1: *WER scores of each speech-to-text model*

3. Text to Text

In the text-to-text task, the goal is to develop a model capable of generating an appropriate response, given a text question as input. Even though context can potentially be offered as input as well, the goal is not to create an information retrieval process. As such, 3 pre-trained large language model with a task of text generation were selected: the GPT-2 [3], Alpaca [4] and the mT0 [5]. The GPT-2 is a well-known text generation model, trained to predict the next word in a sentence on a large corpus of english text. The Alpaca is an instruction-following text generation language model, which in turned makes used of the pre-trained LLaMa language model. Finally, the mT0 is an LLM fine-tuned to generalize well to a variety of different tasks, when given as prompts.

To assess the quality of these models on generating answers, the SQuAD dataset [6], a Question-Answering dataset was used. The Bleu metric was chosen to evaluate each of the LLMs.

For the Alpaca model, the Flan-Alpaca Large variation was chosen.

3.1. Context

The first step was to compare to ability of each individual model when context is given and when it is not. With this in mind, the prompt given as input consists of the context concatenated with the question. The results for one of the questions are shown in table 3. Clearly, adding context leads to a much better performance, for all models except the GPT-2.

3.2. Prompting

Both the Alpaca and mT0 are essentially instruction-following models. It thus expected that they will perform better on our task of question answering when an additional prompt is given. To test this hypothesis, different prompts were added to the input text, before the question is made. 100 samples are randomly selected from the validation split of the SQuAD dataset.

The tested prompts were the following:

- Prompt 1: "Answer this question in less than 10 words. Question: "
- Prompt 2: "Answer the following question: "
- Prompt 3: "Answer this question succinctly: "

The results for the bleu metric are shown in the below table 2, and indicate that Prompt-1 is the most suitable for this task. Additionally, it is also clear that the mT0 performs the best out of all the models. With this in mind, the mT0 with Prompt-1 is the selected combination.

	GPT-2	Alpaca	mT0
Prompt 1	0.00167	0.1594	0.7061
Prompt 2	0.00289	0.1342	0.7050
Prompt 3	0.00178	0.1165	0.6816
No Prompt	0	0.0954	0.5961

Table 2: *Caption*

Finally, different values for the maximum length of the output were tested, and it was found that a maximum length of 25 was sufficiently large for the mT0 to return adequate answers. The Bleu score obtained on the selected pipeline for the first 1000 instances of the validation split of SQuAD was of 0.7197.

”How many BS level degrees are offered in the College of Engineering at Notre Dame?”		
	Context	**No context**
GPT-2	”The number of ”college degrees” is quite small. The number of degrees offered at the University of Connecticut’s College of Law is less than 2%, but the”	”Did we see anyone drop out of engineering degree programs? How many graduate students have started out at a business and have already joined the community? Do graduates of the most prestigious”
Alpaca	Eight B.S. degrees are offered in the College of Engineering at Notre Dame.”	”The College of Engineering at Notre Dame offers a total of 63 BS level degrees.”
mT0	”eight”	”two”

Table 3: *Answers generated for one question in the train split of the SQuAD dataset, with and without context.*

4. Text to Speech

In the final stage, it is necessary to generate an audio of the generated text response. Here, the SpeechT-5 [7]model was used, along with speaker embeddings extracted from the CMU ARCTIC dataset. This task is combined with the previously mentioned ones into an end-to-end dialogue pipeline, and tested on audios for the first 5 questions of the validation split of SQuAD.

5. Conversational Question Answering System

After each of the individual tasks have been implemented, the final step is to combine them into a working dialogue system. Thus, the methods presented in the previous sections are combined.

The CoQA dataset [8] is used to test the developed system, which contains a series of questions of answers in a conversation between two people, along with one story, which gives a context into the questions being made. Specifically, the first sample in the validation split of this dataset is used.

As, ideally, this system should be able to work well when both there is contextual information and when there isn’t, the language understanding and text generation portion is capable of receiving this additional information in the form of text, when it is available. Additionally, the text generation model will also receive the transcript of conversation up to the current question.

The dialogue system was tested on recordings of the first two questions of the first instance of the CoQA validation split, both with and without giving initial context before the conversation starts. As would be expected, the dialogue system performs better, managing to give correct and concise answers.

The main limitation of this system is its difficulty in responding with the correct information when no context is given. In the future, it might be interesting to perform fine-tuning of the LLM on a question-answering dataset, such as SQuAD.

6. References

[1] H. Face, “Transformers documentation,” https://huggingface.co/docs/transformers/, 2021, accessed: June 18, 2023.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.

[4] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, “Stanford alpaca: An instruction-following llama model,” https://github.com/tatsu-lab/stanford_alpaca, 2023.

[5] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, and C. Raffel, “Crosslingual generalization through multitask finetuning,” 2022.

[6] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” *arXiv e-prints*, p. arXiv:1606.05250, 2016.

[7] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, “Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing,” 2022.

[8] S. Reddy, D. Chen, and C. D. Manning, “CoQA: A conversational question answering challenge,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019. [Online]. Available: https://aclanthology.org/Q19-1016