Laboratory Assignment 3 - Dialogue Systems

Group 13
Diogo Miranda - 96190 - diogomiranda26@tecnico.ulisboa.pt
Guilherme Oliveira - 96221 - guilherme.a.oliveira@tecnico.ulisboa.pt

***Abstract*—In this laboratory assignment, we evaluated different models for Automatic Speech Recognition, Language generation and Text-to-speech. Additionally, we studied different strategies to fine-tune these models to our tasks. At last, we developed a conversational spoken question-answering model.**

I. Automatic Speech Recognition

Initially, we collected two audio samples with our voice and used speech recognition models to transcribe the two collected spoken messages. Table I presents the performance of different ASR models on the audio samples using the word error rate (WER) score.

Model	WER Score
OpenAI Whisper Tiny	0.0
Distil Whisper Large v2	0.0
OpenAI Whisper Large-v3	0.0

TABLE I
Performance of ASR models

Due to RAM limitations, we only tested a couple of models. However, we conclude that most Whisper models transcribe accurately the spoken message.

II. Using LLMs for conditional language generation

We used different Large Language Models to do question answering over the TriviaQA dataset, such as Text Generation, Text2Text and Question Answering. Initially, we evaluated the performance of the different implementations on a limited set of 10 questions of the validation set. Table II shows the results obtained. Excluding the Question Answering models which require context, we analyzed for each model the impact of using prompting and context.

Firstly, we conclude that text-generation models have low performance in doing question answering. We observed that GPT2 often continued the question, the given context or the prompting itself and sometimes generated a new completely non-related sentence. The other text-generation models performed more decently and *LaMini-Cerebras* mostly generated answers. However, their score remains very low because they didn't follow the prompt to generate an answer in less than a limited number of words and instead generated many more words than necessary.

Secondly, we conclude that the text-to-text models, excluding the *byt5-small* model which performed poorly, also mostly generate answers to the questions but do not limit the number of words in the sentence. The exception is the *flan-t5-base* that outperforms the other models of its type because most of the times it accurately follows the prompt and limits the length of the answer.

Thirdly, we observe that Question Answering models clearly outperform the other types. We take into account that these models are guided to extract the answer directly from the context and, as a consequence, are overly dependent on the given context.

Overall, we conclude that using context is vital to generate an accurate answer.

Type	Model	Prompting	Context	BLEU2
Text Generation	GPT2	No	No	0
		Yes	No	0
		Yes	Yes	0
	LaMini-GPT-124M	No	No	0
		Yes	No	0
		Yes	Yes	0.020
	LaMini-Neo-125M	No	No	0
		Yes	No	0
		Yes	Yes	0.031
	LaMini-Cerebras-111M	No	No	0
		Yes	No	0
		Yes	Yes	0.047
Text2Text Generation	LaMini-T5-61M	No	No	0
		Yes	No	0
		Yes	Yes	0.051
	LaMini-Flan-T5-77M	No	No	0
		Yes	No	0
		Yes	Yes	0.074
	byt5-small	No	No	0
		Yes	No	0
		Yes	Yes	0.237
	flan-t5-base	No	No	0
		Yes	No	0
		Yes	Yes	0
Question Answering	roberta-base-squad2	Yes	Yes	0.500
	bert-large-uncased	Yes	Yes	0.530
	electra_large_discriminator	Yes	Yes	0.603

TABLE II
Performance of Language Generation Models

Afterward, we compared different prompts to study their impact on the text generated. After experimenting with a different number of instructions to follow and between more straightforward and lengthy prompts, we concluded that it is important to define a well-directed prompt that accurately guides the model to follow a set of instructions.

The models presented in table II were evaluated using a limited part of the context (first 1000 tokens). We also compared the impact of the size of the context on the generated text. On one hand. we observed that decreasing too much the size of the context decreased significantly the performance of the model. On the other hand, we noticed that doubling the size of the context actually led to a slightly worse performance. This might be due to the context being overly extensive and not containing any useful information to the question.

Lastly, we computed the error over the first 1000 examples from the validation set for some selected models using prompting and context and presented it in table III.

Model	BLEU
LaMini-Cerebras-111M	0.022
flan-t5-base	0.149
electra_large_discriminator_squad2_512	0.340

TABLE III
ERROR OVER 1000 EXAMPLES

Once again, we only tested a couple of models due to RAM limitations or limited proprietary permissions. We also tried the versions of the models in table II with more parameters and while most of them exceeded the RAM limitations, a few performed slightly worse on the limited dataset.

Finally, we conclude that the most appropriate model for our dialogue system task is the *flan-t5-base*. We discard Question Answering models because they depend too much on decoding the context information. It is worth mentioning that other models such as MistralAI-Mistral or Meta-LLama-3 might provide better results.

III. Text to Speech

We used the Speech-T5 model to produce speech outputs from the text generated by an LLM. We produced these outputs for the first 5 questions in the TriviaQA dataset and then for 10 audio samples from the SLUE-SQA-5 dataset after connecting the ASR model to the LLM. Lastly, we produced speech-based answers for the first two questions in the validation set in the TriviaQA dataset that were taken as audio samples with our own voice.

Overall, we conclude that the Speech-T5 model accurately generates an audio of the generated text.

IV. Dialogue System

The goal of this task was to integrate previous work and create a conversational spoken question-answering system. To accomplish this, we recorded audio for the questions from one example in the CoQA dataset, used OpenAI Whisper Tiny to transcribe the audio, and then input the transcription into a Text2Text model, *flan-t5-base*, to generate an answer. Finally, we employed SpeechT5 to produce an audio file from the generated answer.

The speech-to-text solution performed exceptionally well, as evidenced by the WER scores, which were nearly zero for each example when compared to the correct questions. The only discrepancies between the transcribed questions and the actual questions were the capitalizations of the character name ”Cotton.” We chose not to convert all questions to lowercase for comparison because, in some instances, the model correctly capitalized the name ”Cotton.” This inconsistency in capitalization highlights the model’s functionality, demonstrating its ability to occasionally recognize proper nouns correctly.

For the text generation part, we devised various texts related to the context provided to the model. We tested multiple scenarios: with no context, with only the context given by the dataset, with the dataset context and the previous questions and answers, and finally, for the last test, we added a prompt to help the model understand the problem. The results are shown in the table below.

Scenario	Average BLEU Score
No Context	0.00
Dataset Context Only	0.34
Dataset Context + Previous Q & A	0.27
Dataset Context + Previous Q & A + Prompt	0.34

TABLE IV
AVERAGE BLEU SCORES FOR DIFFERENT CONTEXT SCENARIOS

The BLEU score was calculated with a maximum order of 2, comparing the correct answers from the dataset with the generated answers. For dataset answers that contained only one word, we used a BLEU score with an order of 1. This adjustment was made because, in cases where the model generated the same one-word answer as the dataset, the BLEU score would otherwise be 0 instead of 1.

Based on the results in the table IV, we observed that the best scores were achieved using either the dataset context only or the dataset context combined with the previous questions, answers, and a prompt. The prompt was necessary because using only the original context along with the previous questions and answers confused the model, causing it to repeat answers to previously answered questions. Despite the inclusion of the prompt, this issue still occurred, although less times.

Analyzing the actual answers, one of the main issues was that the generated responses for many questions were too long. We attempted to instruct the model in the prompt to keep the answers under a certain number of words, but this strategy only worked for certain responses, and we still received many overly lengthy answers.

We can conclude that we successfully developed a conversational spoken question-answering model. However, the results could be improved by further experimenting with different models and prompts.

V. Conclusion

In this laboratory assignment, we learned how to implement several models for ASR, Text Generation and TTS to develop a dialogue system. We understood the value of using context to generate text, as well as the importance of defining an accurate prompt to guide a model to follow a set of instructions.