Lab 3 SLP 2022/2023, Group 3

95565 Duarte Almeida, 95618 Leonor Barreiros

Abstract

The development of dialogue systems has been an active research problem for many years. Recently, with the development of large language models, we have been seeing a lot more chat-based applications. In this project, we propose a dialogue system for chit-chat: a natural conversation that's as human-like as possible.

1. Experimental Setup

Our research followed three typical steps in the scientific method: getting to know the field and formulating our goal, designing the experiments to test our model, and implementing and evaluating the final system.

The goal of our research is to develop an end-to-end dialogue system [1] which can seamlessly carry a dialogue with a human. To achieve this, we propose a model we will detail in section 2.

Our experiments were conducted using the **S**tanford **Qu**estion **A**nswering **D**ataset [1], which consists of a set of questions and respective answers, with a context, where the answer to each question is within the context. When selecting between alternative approaches, we used the 200 first samples of the validation split; the final results were calculated using the 1000 first samples of the same split. On simpler tasks, namely Automatic Speech Recognition, experiments were conducted using self-recorded audios.

Finally, we will describe the implementation and evaluation of our system in the upcoming sections.

2. Approaches

Our dialogue system is composed of three modules (note that we don't follow the modular structure of task-oriented dialogue systems because ours is an end-to-end, chit-chat based system): Automatic Speech Recognition (ASR), conditional language generation, and text-to-speech conversion.

We determined the best architecture for each one independently, and will analyze each one's options next.

2.1. Automatic Speech Recognition

Automatic Speech Recognition (ASR) is a **sequence-to-sequence** problem, where we classify a speech signal with a sequence of words (the words said in the signal).

This task can be solved in a black-box manner by choosing among many models provided in the HuggingFace library. In fact, all the tasks in our problem can be solved resorting primarily to this strategy.

Whisper [2] is a pre-trained model for ASR that learned to perform this task in a supervised manner. We evaluated 3 variants of this model: tiny (39M parameters), small (244M parameters) and medium (769M parameters). On the two pre-recorded samples we evaluated, the small and medium models behaved the same way, however, to be conservative and contemplate more intricate signals, we opted for the medium model.

SpeechT5 [3] is a fine-tuned model (from T5) for ASR. It performed worse than all versions of whisper, which we think is due to it having been trained on unlabeled data.

2.2. Conditional Language Generation

Conditional Language Generation is the most crucial module of our system. We searched for a Large Language Model (LLM) that worked well under the following conditions:

- Perform correct question answering when the answer to the question is within the context (the case for SQuAD), but in a **generative** way. For example, if we had used an extractive model, examples of the sort "CONTEXT: *the dog is black*, QUESTION: *is the dog white?*" could have no answer
- Perform a human-like question answering when the questions are simple chit-chat. This latter is more difficult to accomplish, especially when considering our limited computational resources (we can't use a natural model like ChatGPT, for instance)
- Is fast to answer. If we want our system to be human-like, we cannot leave the interlocutor waiting for a reply for long, or else the conversation won't flow

As such, we evaluated a lot of different alternatives to come up with the model that fulfilled these properties best.

Our first approach consisted of prompting GPT-2 with the context of the question and "*The shortest and simplest answer possible is:*". We also experimented using this approach, but with the addition of providing **extra context** to the prompt, which was determined by extracting information via IR techniques [4]. More specifically, we recurred to all-MiniLM-L6-v2 [5], a sentence embedding model that is able to assign a 384 dimensional vector representation to each context-question pair in the training dataset. At test time, given a context-question pair, the most similar context-question pair present in the training set according to cosine similarity was appended to the prompt so as to perform some sort of knowledge augmentation.

Furthermore, we conducted the same experiments but with an Alpaca-based model.

Then, we experimented with two different fine-tuned models for QA, prompted with the context besides the question, both based on T5 [6] [7]. Finally, motivated by the fact that there

[1] Although our system has 3 independent modules, we consider it an end-to-end system because the ASR and TTS modules are typically separate, and it's the LLM that handles language understanding and dialogue management.

may not be a context, and that we want a seamless conversation between machine and human, we also experimented with DialoGPT [8].

Our final model uses a fine-tuned T5 for QA in the case that there is a conversational context, and DialoGPT if there isn't. We will discuss our decisions and results in Section 3.

2.3. Text-to-Speech Conversion

For the final module, we used SpeechT5. Note that we also experimented using it for ASR; that's because this model is capable of performing three types of tasks: speech-to-text (e.g., ASR), text-to-speech (what we used it for), and speech to speech (e.g., speech enhancement) [9].

3. Results and Discussion

For the first intermediate task, **ASR**, the selected model (Whisper medium) achieved an average Word Error Rate of around 0.07. Whisper small had a larger WER, 0.29, and T5 performed the worse, with a WER of around 0.54.

The results for the second intermediate task, **Language Generation**, are in Table 1. It was expected that the first fine-tuned T5 would perform the best, since it used the same dataset for evaluation. We also highlight the fact that the addition of in-context examples only led to marginal improvements in the results and, since it took a long time to run, we opted not to use it. Finally, although DialoGPT had a bad score, we decided to use it for the case when there's no conversational context, since it gives more human-like answers. Note however that these results are somehow misleading in assessing the conversational capabilities of these models, since in many occasions they generate the right answers but not in the most concise manner, even when prompting them in that direction. The poor performance is then explained by the fact the BLEU penalizes predictions longer than the reference text, being agnostic to semantic aspects of the generated text.

Model	BLEU
GPT-2	0.0
GPT-2 with IR	0.0
Alpaca	0.25
Alpaca with IR	0.26
T5 for QA 1	0.53
T5 for QA 2	0.16
DialoGPT	0.0

Table 1: *Results of different approaches for conditioned language generation.*

Since we used a single approach for the final intermediate task, **TTS**, we didn't measure the results in a quantitative manner.

4. Concluding Remarks and Future Work

The key component in developing a dialogue system is the one where we generate a written response given the posed question, i.e., conditional language generation. However, it is also the one, and especially taking into consideration our limited computational and time resources, where it was the most difficult to achieve a good performance. Only with the recent developments of LLM has it been possible to achieve satisfactory results.

Moreover, when developing a system that's meant to be used in real time, there are more requirements than quantitative metrics that we need to take into consideration. For example, we took great concern in ensuring that we used a model that **(1)** replied fast, as to not break the flow of the conversation, and **(2)** have as human-like answers as possible. Because of that, we ended up trading-off a model with a bad BLEU score with the fact that it gave very fast, very human-like responses (DialoGPT).

In the future, we intend to fine-tune the language model. This is due to two reasons: firstly, because even when there is a context, sometimes the LM provides no answer, and in that case we could try to extract the response or generate it with another model; secondly, because the model we used for when there is no context is very limited, and many times gets stuck in repetitive answers, so we also have some improvements to make in that scenario.

5. References

[1] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv preprint arXiv:1806.03822*, 2018.

[2] S. Gandhi, "Fine-tune whisper for multilingual asr with huggingface transformers," https://huggingface.co/blog/fine-tune-whisper, accessed: 2023-06-10.

[3] 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," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 5723–5738.

[4] B. Martins, "Dialogue systems," fenix.tecnico.ulisboa.pt/downloadFile/282093452111441/Dialogue_Systems_2.pdf, accessed: 2023-06-16.

[5] "all-minilm-l6-v2," https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, accessed: 2023-06-17.

[6] G. N. Christian Di Maio, "T5 for generative question answering," https://huggingface.co/MaRiOrOsSi/t5-base-finetuned-question-answering, accessed: 2023-06-10.

[7] "Question answering generative," https://huggingface.co/consciousAI/question-answering-generative-t5-v1-base-s-q-c, accessed: 2023-06-10.

[8] "A state-of-the-art large-scale pretrained response generation model (dialogpt)," https://huggingface.co/microsoft/DialoGPT-medium?text=Hi, accessed: 2023-06-15.

[9] M. Hollemans, "Speech synthesis, recognition, and more with speecht5," https://huggingface.co/blog/speecht5, accessed: 2023-06-15.