

The Copperbelt University

School of

Information and Communication Technology

**Question-Answering System using GPT-2**

**Group members**: Frank Lembalemba, Chota Mulenga, Richard Simbula and Charles Chisale

**Student Numbers**: 20148331, 20167136, 20153999,20152155,

**Program**: Computer Science

**Course**: Artificial Intelligence

**Project Overview**

This documentation outlines the development and implementation of a Question-Answering (Q&A) system leveraging the GPT-2 language model. The system allows users to input questions about specific individuals and generates corresponding answers based on pre-defined knowledge.

**Components**

1. **Data Format:**
   * Questions and answers are structured using special tokens (<Q> for questions and <A> for answers) to facilitate parsing and understanding by the model.

Example:

plaintext

Copy code

<Q> Who is Richard Simbula? </Q>

<A> Mr Richard Simbula is a final year computer science student. </A>

1. **Model Selection:**
   * The GPT-2 (Generative Pre-trained Transformer 2) model is selected for its ability to generate coherent text based on context and input questions.
2. **Training Data Preparation:**
   * The training dataset comprises curated question-answer pairs formatted with special tokens to train the GPT-2 model to associate questions with their appropriate answers.
3. **Model Fine-tuning:**
   * Fine-tuning of the GPT-2 model is performed using the formatted training dataset to optimize its ability to generate accurate and contextually relevant answers.
4. **Deployment:**
   * Gradio is utilized for deploying the Q&A system, providing a user-friendly interface where users can input questions and receive answers generated by the fine-tuned GPT-2 model in real-time.

**Methodology**

**Data Collection and Preprocessing**

* **Data Collection:** Information about individuals (e.g., Richard Simbula, Frank Lembalemba) is gathered from reliable sources and formatted into structured question-answer pairs.
* **Data Formatting:** Each individual profile is formatted using special tokens (<Q> and <A>) to distinguish questions from answers and ensure compatibility with the model.

**Model Fine-tuning**

* **Model Selection:** The GPT-2 model is chosen for its proven effectiveness in natural language processing tasks, specifically text generation.
* **Fine-tuning Process:**
  + The pre-trained GPT-2 model is loaded and initialized.
  + The model is fine-tuned on the formatted question-answer pairs to adapt its parameters to the specific task of generating accurate answers to input questions.

**Implementation**

1. **Data Preparation:**
   * **Structured Formatting:** Individual profiles are transformed into question-answer pairs using special tokens for clarity and model compatibility.
   * **Dataset Creation:** The formatted data is compiled into a single training dataset file (train.txt) for efficient model training.
2. **Model Training:**
   * **Tokenization:** The GPT-2 tokenizer is utilized to tokenize and preprocess the training dataset.
   * **Model Initialization:** The GPT-2 model is initialized and fine-tuned using the formatted dataset to optimize for question-answering tasks.
   * **Training Configuration:** Hyperparameters such as batch size, number of epochs, and learning rate are set for optimal training performance.
3. **Deployment Using Gradio:**
   * **Interface Design:** A Gradio interface is designed to interact with the fine-tuned GPT-2 model.
   * **User Interaction:** Users input questions through the Gradio interface, which processes the input using the model and returns generated answers in real-time.

**Example Usage**

* **User Input:** "Who is Richard Simbula?"
* **System Output:** "Mr Richard Simbula is a final year computer science student at the Copperbelt University."

**Conclusion**

This project demonstrates the successful implementation of a Q&A system using the GPT-2 model, integrating structured data formatting, model fine-tuning, and user interface design via Gradio. The methodology ensures robust data preparation and model adaptation, while the implementation showcases effective deployment for user interaction.

**References**

1. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI. Retrieved from <https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf>
2. Hugging Face. (2024). Transformers Library Documentation. Retrieved from https://huggingface.co/transformers/
3. Gradio. (2024). Gradio Documentation. Retrieved from https://gradio.app/docs/