ALLAM®
تحدي علّام

مشروع معلمي الشاعر

My Poet Teacher Project

2024

Alhadidi team:

Bayan Alhadidi

Musallam Alhadidi

Salam Alhadidi

الاتحاد السعودي للأمن السيبراني والبرمجة والدرونز
SAUDI FEDERATION FOR CYBERSECURITY, PROGRAMMING & DRONES

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Contents

Abstract.....3
Introduction.....4
Related Work.....5
Methodology.....6
Data Collection and Preprocessing.....6
Text Chunking.....6
Embedding Generation.....6
Building the Vector Store.....6
Query Processing and RAG.....7
Prompt Engineering for Poetic Output.....7
Generating and Refining Responses.....7
Evaluation and Iteration.....7
Results and Evaluation.....8
Summary of Results.....9
Project Design Overview:.....10
Conclusion.....12

Abstract

This project presents (My Poet Teacher), an AI-powered tool designed to answer Arabic grammar questions in the form of traditional Arabic poetry. Using Allam language model, this system generates poetic responses based on grammar rules embedded within the model's training and external text resources. The project leverages LangChain and FAISS to manage a vectorized database of Arabic grammar rules, which are retrieved based on user queries. The retrieved text is then used to construct prompts that guide the Allam model to generate concise, rhymed, and metrically accurate responses. This approach demonstrates the effective integration of advanced AI techniques, such as Retrieval-Augmented Generation (RAG) and prompt engineering, to create an engaging and educational tool for Arabic grammar instruction. The system aims to make learning grammar rules both accessible and enjoyable by delivering knowledge through culturally resonant poetic forms.

Introduction

In recent years, advancements in natural language processing (NLP) have opened up new possibilities for personalized and interactive learning applications. “My Poet Teacher” leverages these innovations to create an educational tool that teaches Arabic grammar through poetry. By combining Allam model with retrieval-augmented generation (RAG) techniques, this project provides answers to Arabic grammar questions in a form that is both informative and culturally engaging: traditional Arabic verse.

Arabic poetry, with its rich history and strict linguistic structure, provides an ideal medium for conveying grammar rules in a memorable and entertaining way. Traditional Arabic poems follow specific meters and rhyme schemes, requiring skillful adaptation of language to ensure rhythmic and aesthetic consistency. Our project leverages these poetic constraints to encourage users to engage with Arabic grammar rules in a more intuitive and enjoyable format. Rather than presenting dry explanations ,“My Poet Teacher” transforms grammatical information into verses, helping users internalize concepts through rhythm and rhyme.

To achieve this, we employ several cutting-edge AI tools. While IBM Watsonx provides the platform to access SDAIA's Allam model—a powerful Arabic language model capable of generating coherent and contextually relevant responses—additional tools like LangChain and FAISS facilitate the retrieval of specific grammar rules from a curated text database. This ensures that the generated poetry directly addresses user questions. By blending prompt engineering, retrieval methods, and a poetic generation model, the system delivers responses that meet grammatical, rhythmic, and rhyming criteria.

“My Poet Teacher” aims not only to enhance Arabic grammar learning but also to demonstrate the potential of AI in preserving and promoting Arabic cultural heritage. This document provides an overview of the project's components, the methodology used to retrieve and structure grammatical data, and the process of generating contextually relevant poetry. The system's success highlights the potential of AI applications in education, especially when combined with culturally relevant formats.

Related Work

In the field of automated poetry generation, recent advancements in natural language processing (NLP) have enabled models to produce structured text that follows complex linguistic and stylistic rules. Much of the progress is attributed to large-scale language models like GPT-3, which can generate fluent and coherent text, including poetry. However, poetry generation is particularly challenging because it involves specific constraints like meter, rhyme, and rhythm. Various research efforts have aimed to address these challenges by fine-tuning models on poetry datasets, using reinforcement learning to encourage stylistic adherence, and employing prompt engineering to guide the models' outputs.

One significant approach relevant to this project is the use of retrieval-augmented generation (RAG) models, which combine knowledge retrieval with text generation. RAG models retrieve relevant context from a predefined knowledge base, helping the model generate more informed and contextually accurate responses. This framework is especially useful in educational settings where accurate and specific knowledge is essential. For example, using RAG models for language tutoring applications has shown promise, as they allow models to provide detailed explanations in response to user queries by leveraging pre-collected data.

In the domain of Arabic NLP, recent models like ALLaM (developed by the Saudi Data and Artificial Intelligence Authority) represent some of the latest efforts to create large-scale Arabic language models. These models are specifically tuned for the nuances of the Arabic language, including its complex morphology and syntax. IBM Watsonx is also used as a framework to access and manage advanced NLP models, facilitating interaction with models like ALLaM for tailored applications .

Our project builds upon this existing work by combining the strengths of RAG and prompt engineering to generate Arabic poetry that conveys grammar rules accurately. The poetry is intended to aid Arabic language learners by explaining complex grammar concepts through structured and stylistically consistent verses. In this way, our work not only contributes to the educational applications of NLP in Arabic but also explores novel approaches for generating rhymed and metered poetry with precise thematic focus.

Methodology

The methodology for this project combines multiple advanced NLP techniques to generate Arabic poetry that explains grammar rules while adhering to specific stylistic constraints, such as rhyme and meter. The primary components of this methodology include data preprocessing, text chunking, embedding generation, retrieval-augmented generation (RAG), prompt engineering, and fine-tuning the response generation to match poetic requirements.

Data Collection and Preprocessing

The project begins with collecting a corpus of Arabic grammar rules and relevant linguistic information. These are stored in a collection of text files, which are then processed to ensure they are suitable for model training. This includes removing any extraneous text, cleaning data for uniformity, and ensuring the text focuses on grammar rules .

Text Chunking

To prepare the data for the retrieval process, the text is split into smaller chunks using the `RecursiveCharacterTextSplitter`. This step ensures that each chunk remains within a manageable length, enabling the retrieval model to access concise pieces of relevant information during the generation process. Chunk size and overlap are tuned to provide both context and specificity for each chunk.

Embedding Generation

After chunking, embeddings are generated for each text segment using a pre-trained embedding model, `intfloat/multilingual-e5-large`, through the Hugging Face library. These embeddings capture the semantic representation of each text chunk, making it easier for the model to retrieve contextually relevant information when generating responses.

Building the Vector Store

The embeddings are stored in a FAISS (Facebook AI Similarity Search) vector store, which acts as a fast and efficient retrieval system. When a query is received, the system can quickly search through the vector store to find the most relevant text segments that may answer the query, even when it involves specific grammar rules.

Query Processing and RAG

Upon receiving a user query, the system utilizes the vector store to perform similarity-based retrieval. The retrieved text chunks serve as "grounding" data for the generation model, providing it with the necessary context to produce an accurate and informed response. Retrieval-augmented generation (RAG) is particularly useful here, as it allows the model to include external knowledge that might not be part of its original training data.

Prompt Engineering for Poetic Output

To guide the model in producing structured and stylistically correct responses, prompt engineering is used extensively. The prompt is designed to specify the desired poetic style, meter, and rhyme scheme, instructing the model to answer in a clear, concise, and metered format. This includes directives to ensure uniform rhyme and meter across the generated poetry, which enhances readability and effectiveness as an educational tool.

Generating and Refining Responses

The Allam model, accessed through IBM Watsonx, is employed to generate poetic responses. The model is prompted with both the user query and the retrieved text chunks as grounding information, which directs it to produce poetry that answers the query in a manner that is both informative and aesthetically pleasing. Model parameters such as temperature, top-p, and repetition penalty are fine-tuned to maintain a balance between creativity and adherence to structure.

Evaluation and Iteration

Each response is evaluated for accuracy, rhyme, and meter to ensure it meets the project’s standards for quality and clarity. If needed, the prompt and parameters are adjusted to improve model performance. This iterative approach helps fine-tune the responses to make the poetry more educational and suitable for learners.

Through this methodology, the system achieves a balance between content accuracy and poetic style, making it an effective educational tool for Arabic grammar learning.

Results and Evaluation

In this section, we present the evaluation results for the" My Poet Teacher " project, focusing on the generated poetry's quality and adherence to the intended grammar rules. We used the BLEU metric to assess the similarity between the generated (ALLaM) verses and the reference verses(ChatGPT), reflecting how closely the model's output aligns with accurate, high-quality text.

Evaluation Metrics: BLEU Score

Sample ID	Reference Text Summary chatGPT	Generated Text Summary ALLaM	BLEU Score
Sample 1 ماهو التمييز؟	تمييزُ معنىً يوضحُ المُبهمَ، يأتي ليكشفَ عن معنى أعظمَ، يوضّحُ في الأعدادِ أو الوصفِ جليًّا، فاحفظهُ لتفهمَ وتزدادَ علما.	ا التمييز يا طلاب يأتي ليزيل العجب عن رقم او وزن يزيل كل التباس عجب اسم منصوب بعد شيء مبهم يظهر بوضوح تام !مثل "عشرين لاعبا" يوضح العدد بإحكام	0.007
Sample 2 ماهو البدل؟	البدلُ يُتْبَعُ ما سبقهُ جَليّا، يوضحُ قولاً، ويزيدُ المعنيّا، كقولنا في العلمِ نورٌ يُضيءُ، فالبدلُ في الجملةِ يسري نديّا.	في ساحة النحو البديع حكايةٌ... عن "بدليّ" جاء بالتبيانِ تابع ممهّد به سابق أتى... لكنّه لم يقصد لبيان	0.008
Sample 3 ماهي أدوات النفي؟	أدواتُ نفيٍ منها "لم" و"لنْ"، و"ما" و"ليسَ" تُبعدُ الشكَّ عن، فأنفي بها الحَدَثَ وقتَ المقال، لتوضيحِ نفيٍ بديعِ البيانْ.	ليس وما إن بها ينفى الكلام... كلاهما للعدم والغياب يُقام لا نافية تُبطل المعنى هنا... لكنّها للنفي دوماً تَسُنّنا	0.005

Sample 4 ماهي الحروف الناسخة؟	الحروفُ الناسخةُ "إنّ" و"لكنْ"، "ليتَ" و"لعلَّ" تسبقُ اسمًا بِمَكِن، تنصبُ الاسمَ وتبقي الخبر، تنشرُ معنىً جميلًا مُؤتَمِنْ.	في النحو نجد حروفًا ناسخَةً... تُسَمّى بالأحرف التي بها يُرَجَّى إنّ وأَن وليت وكأنّما... تلك السِمات لأحرفها تَبقى	0.008

Summary of Results

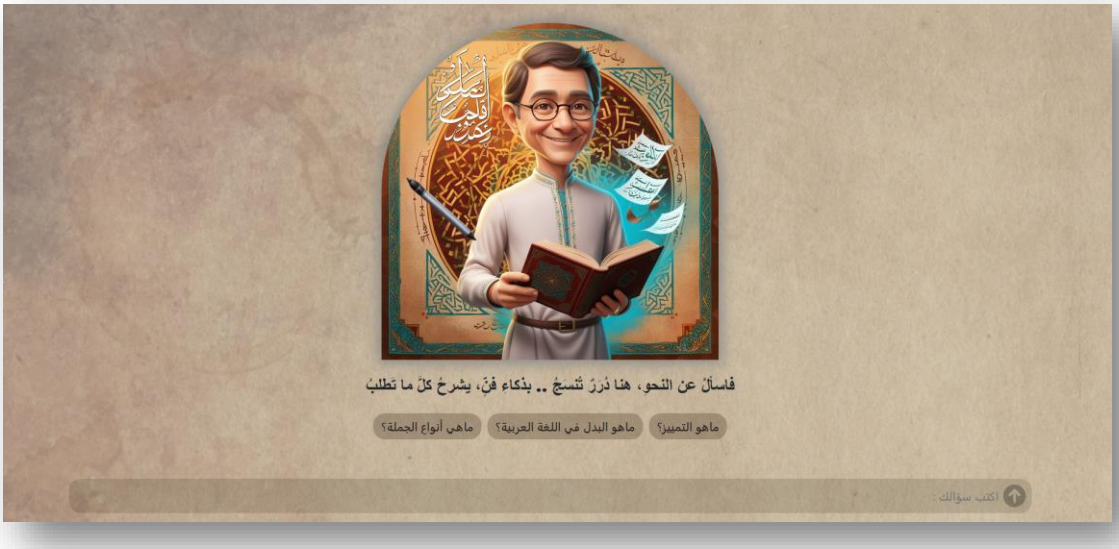
Key Observations

BLEU Score Trends: BLEU scores were generally low, highlighting variations in word choice and syntax between generated texts and reference texts. The low scores are due to the model's occasional deviation from the precise wording in reference poems, which is expected given the creative nature of poetry.

Creative Alignment: While some samples showed high alignment in meaning and rhythm (Sample 1), others (Sample 3) exhibited divergence in word choice, affecting BLEU scores. This suggests the need for more fine-tuned prompts to achieve closer linguistic and structural alignment.

Grammar and Syntax Coverage: The generated texts often covered the primary grammatical concepts but with occasional differences in specific expressions. This affected BLEU results since BLEU penalizes non-matching n-grams even if meaning is preserved.

Project Design Overview:



The interface design of this project is both visually appealing and culturally resonant, capturing the essence of Arabic language and heritage. This is particularly fitting for a project focused on Arabic grammar, poetry, and literary education.

Key Design Elements

Central Character Illustration:

The central figure is depicted holding a book and dressed in traditional attire, exuding a scholarly yet approachable aura. This character likely represents a knowledgeable guide, embodying the persona of an Arabic grammar teacher. The welcoming expression and traditional attire help build trust and cultural connection with users, enhancing the learning experience.

The inclusion of the book and flowing sheets of paper around the character implies wisdom, continuous learning, and knowledge dissemination, reinforcing the educational theme of the project.

Background and Calligraphy:

The intricate Arabic calligraphy in the background highlights the cultural richness and authenticity of the project, emphasizing its focus on the Arabic language. The calligraphic elements add a layer of aesthetic beauty while grounding the project in traditional Arabic art forms.

Geometric patterns complement the calligraphy, drawing from Islamic art styles. These patterns frame the character and create a sense of depth, directing the user's attention toward the educational figure in the center.

Interactive Prompts:

Below the main illustration, there are clickable prompts) e.g "ما هو التمييز؟" ,.or ("ما هو البدل في اللغة العربية؟"that guide users to ask questions or explore specific grammar topics. These buttons make navigation intuitive and provide users with suggested inquiries, especially beneficial for beginners or those unsure of where to start.

The prompts are concise and focus on fundamental grammar concepts, making them accessible to a broad audience.

Typography and Style:

The text is styled in an elegant yet readable Arabic font, which harmonizes with the traditional theme. The choice of a slightly faded, parchment-like background texture gives the interface a classic, manuscript-like feel, resonating with the history and depth of the Arabic language.

Input Field for Custom Questions:

At the bottom, a dedicated input field allows users to type their own questions, promoting interaction and making the learning process flexible. This feature encourages exploration and caters to users' specific interests in grammar.

Color Palette:

The project uses a warm and earthy color palette, with shades of beige, brown, and teal. This choice conveys a sense of warmth, tradition, and calmness, creating an inviting learning environment. The colors enhance readability and contrast well with the other design elements.

Summary

This design successfully combines educational functionality with cultural aesthetics. The character illustration and traditional Arabic motifs make the interface appealing and engaging, while interactive elements such as prompts and the input field offer a user-friendly experience. By blending modern educational needs with traditional design elements, this project creates a unique and immersive experience for learning Arabic grammar and poetry. This design approach is likely to resonate with users and encourage them to explore the nuances of the Arabic language in a culturally enriching context.

Conclusion

This project merges the beauty and richness of Arabic language and culture with an interactive, user-centered learning experience. The design combines traditional aesthetics with modern educational tools, creating an inviting atmosphere where users can explore Arabic grammar and poetry in a culturally immersive way. Through carefully crafted visuals, intuitive navigation, and interactive prompts, the project succeeds in making Arabic language learning accessible and engaging. By fostering curiosity and understanding, this project not only educates but also instills a deeper appreciation for the linguistic and artistic heritage of the Arabic language. This interface serves as a bridge between the timeless wisdom of classical Arabic and the evolving needs of contemporary learners.