Batch ID : CIN238

INFORMATION EXTRACTION USING LARGE LANGUAGE MODELS

RA2011026010122
Dibyajyoti Ganguly

RA2011026010132
Tarun Negi

Guide : Dr. Maivizhi R
Assistant Professor, CINTEL

ABSTRACT

- Efficiently extracting information from documents remains a critical challenge in various domains. As we witness an exponential growth in online sources, the need for efficient and intelligent tools to distill and comprehend this wealth of information becomes more pronounced. The traditional methods of research often fall short in managing the sheer volume of data available, leading to information overload and reduced productivity. Our proposed work leverages the powerful combination of Langchain and OpenAI.

- This innovative approach is designed to streamline research by allowing users to input documents and extracting pertinent information based on customized prompts. This enables users to provide specific prompts, tailoring the extraction process to focus on particular aspects of the news articles. The experimental results demonstrate the effectiveness of our approach in extracting the information from PDFs.

INTRODUCTION

- In the dynamic landscape of information dissemination, staying well-informed is crucial, and research plays a pivotal role in understanding and interpreting current events. With the proliferation of online news sources, researchers, journalists, and enthusiasts often grapple with the overwhelming volume of information.

- The tool aims to simplify this process, where users can input news articles and receive targeted and succinct summaries based on their specified prompts. It extracts relevant facts but also distills the essence of documents, providing users with a comprehensive yet concise overview of the content. The capacity to effectively extract pertinent responses from enormous repositories of textual data is critical in today's information-rich environment

MOTIVATION

- The motivation behind developing the PDF Query Tool stems from the increasing complexity and volume of information in the digital age. As we witness an exponential growth in online news sources, the need for efficient and intelligent tools to distill and comprehend this wealth of information becomes more pronounced.

- The traditional methods of news research often fall short in managing the sheer volume of data available, leading to information overload and reduced productivity. This project is motivated by a desire to bridge this gap by harnessing the capabilities of Langchain and OpenAI, creating a solution that not only simplifies the research process but also enhances its precision and relevance.

CHALLENGES AND LIMITATIONS IN EXISTING SYSTEMS

- Natural language is often ambiguous, and understanding the context of a news article can be challenging. Extracting the intended meaning accurately, especially in cases of sarcasm, irony, or nuanced language, remains a difficulty. Existing systems struggle to grasp the full context of an article, leading to potential inaccuracies in information extraction.
- Tools like chatGPT may somewhat be able to accurately extract information from the news articles but it is a tedious process to feed entire articles everytime for information extraction. Also, a proper knowledge base is required along with a tool to bypass chatGPT’s word limit of 3000 words.

OBJECTIVES

- Developing a system that accurately extracts key information from documents, including facts, quotes, and contextual details, to ensure the reliability of the extracted content.

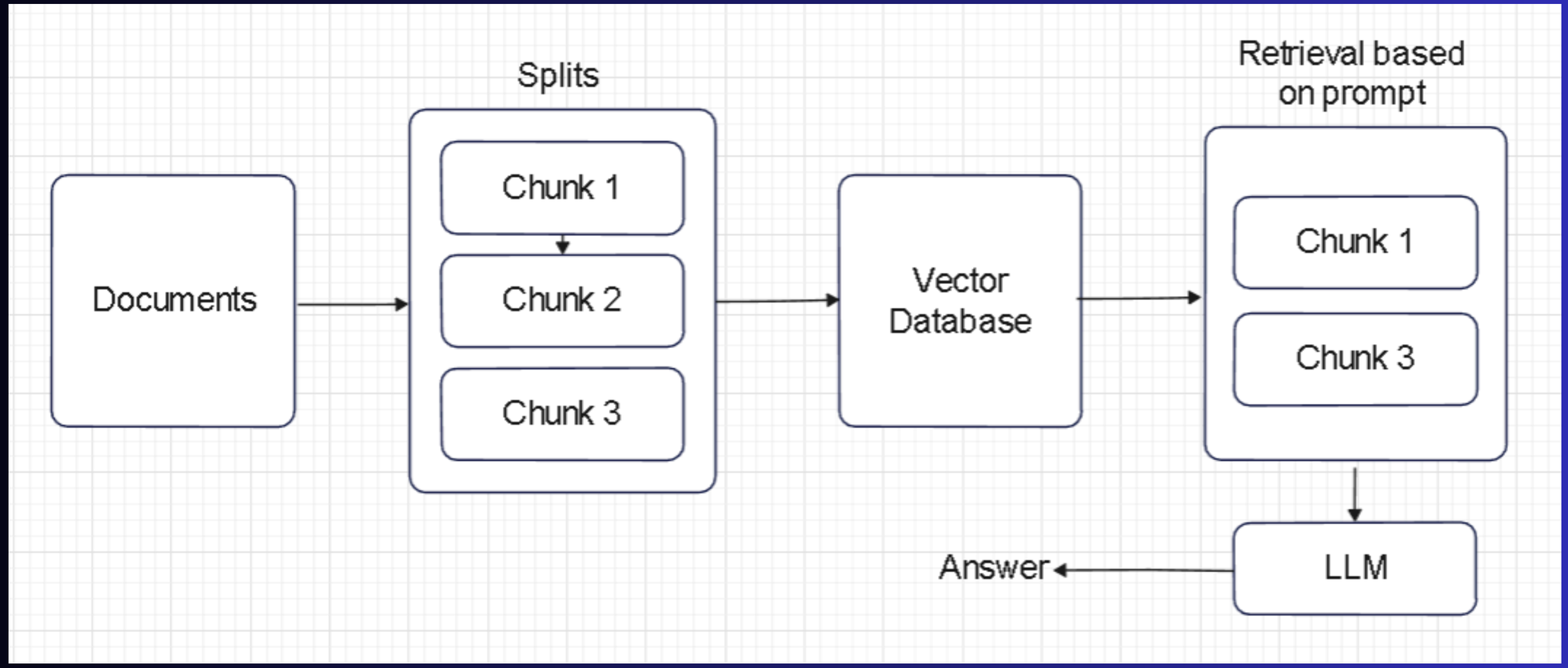
- Enhancing the system's ability to understand and interpret the context of articles, addressing challenges related to ambiguous language and nuanced expressions.

- Providing users with the ability to input customized prompts, tailoring the information extraction process to meet individual or specific organizational needs.

SCOPE AND APPLICATION

- Journalists and reporters can use the tool to quickly extract key information from multiple news articles, aiding in the research phase and enabling them to stay on top of developing stories.
- Businesses can leverage the tool for competitive intelligence, extracting and summarizing information from news articles relevant to their industry and competitors.
- It can be used for equity research analysis to stay updated on latest stock trends.
- Educational institutions can incorporate the tool to teach students about media literacy, helping them critically analyze news articles and understand how information is extracted.

ARCHITECTURE



MODULE DESCRIPTION

1) Input Module : This module is responsible for taking user input, which includes the article in the form of a PDF and any custom prompts provided by the user.

2) Information Extraction Module : This module is at the core of the system, focusing on extracting relevant information from the article based on the document and user prompts.

RESULT ANALYSIS

Chunk Size : With a chunk size of 1000, the system achieved higher precision but lower recall compared to a chunk size of 500. This suggests that larger chunk sizes leads to more accurate answers.

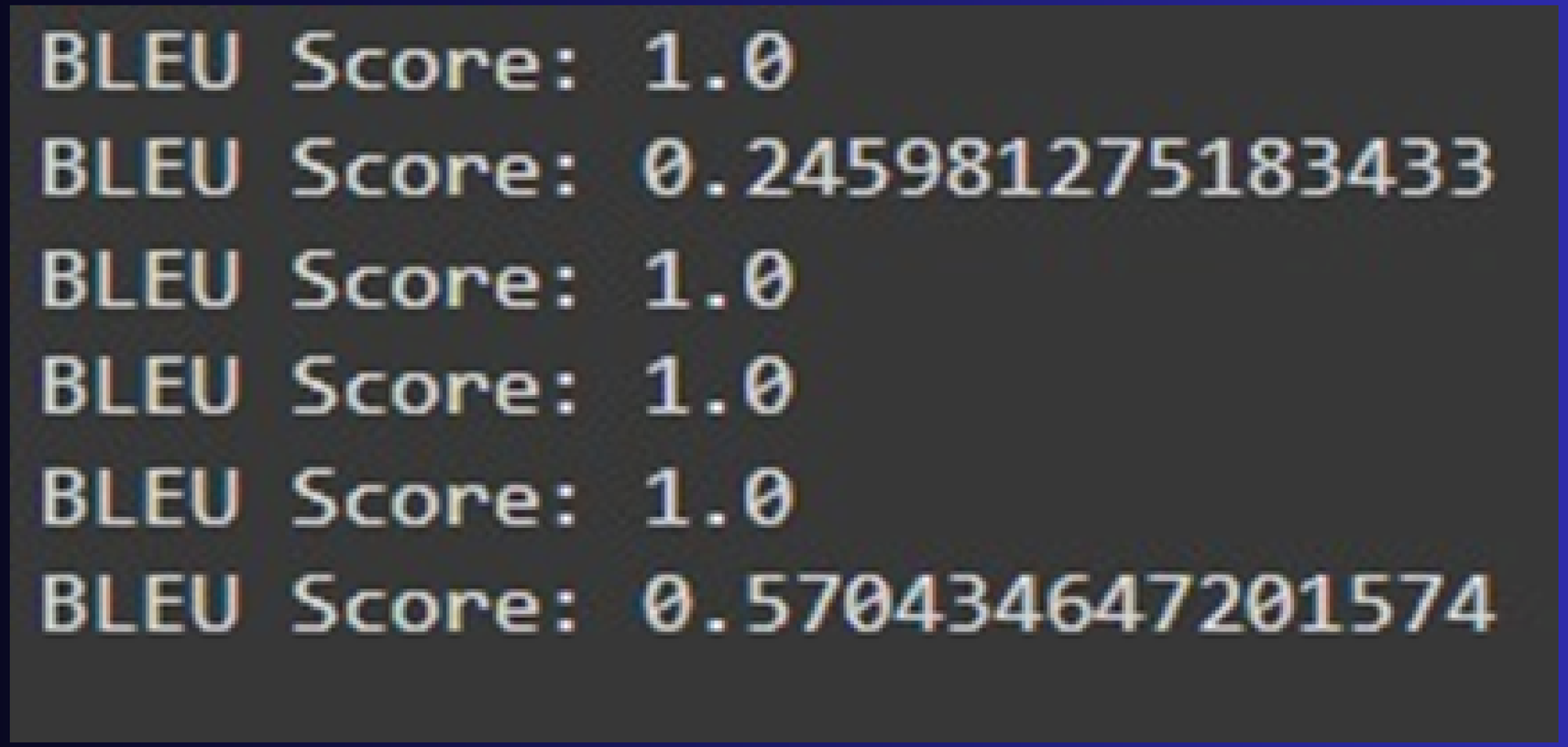
BLEU score : We have also used BLEU (BiLingual Evaluation Understudy) scores as a metric to compare the LLM's. The BLEU score gives an output score between 0 and 1. A BLEU score of 1 depicts that the sentence perfectly matches one of the reference sentences.

The OpenAI LLM gives an average BLEU score of 0.8

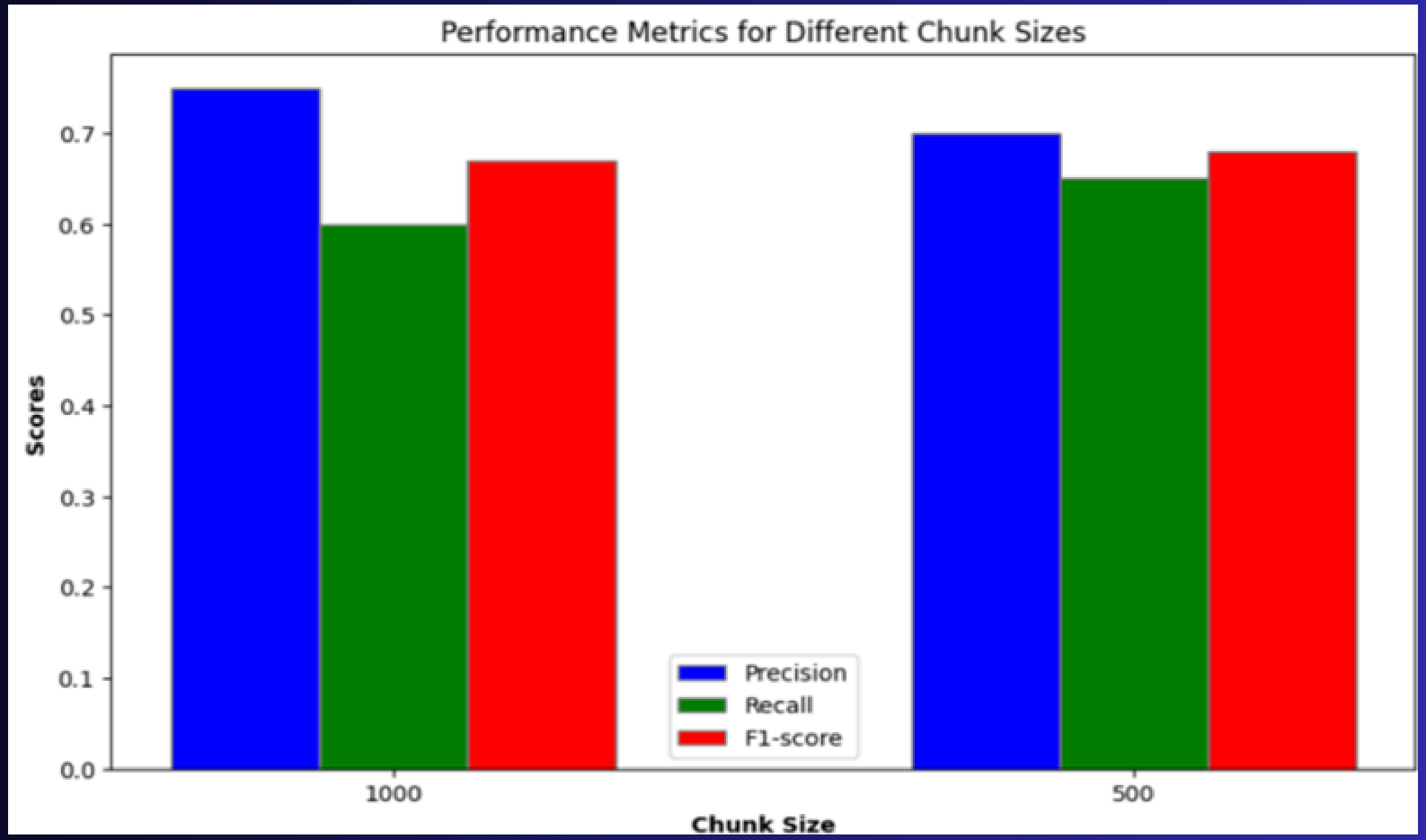
The adopted LLM's BLEU score corresponding to multiple query and responses is shown below. Higher BLEU score indicates better performance in generating translations that match human references.

```
# Reference text for each query
reference_texts = [
    "Authors of the article are John Doe and Jane Smith.",
    "The exact aggregation level aggregation algorithm from the document is Algorithm 1.",
    "Aggregation algorithm with all mathematical complexities includes multiple steps such as matrix multiplication and sorting.",
    "Algorithm 2 for aggregator level aggregation can be found in the document.",
    "The content inside Algorithm 2 for aggregator level aggregation includes steps for data preprocessing and feature extraction.",
    "Performance evaluation summary: The model achieved an accuracy of 90% on the test dataset."
]
```

```
# Generated text for each query
generated_texts = [
    "Authors of the article are John Doe and Jane Smith.",
    "Algorithm 1 is the exact aggregation level aggregation algorithm.",
    "Aggregation algorithm with all mathematical complexities includes multiple steps such as matrix multiplication and sorting.",
    "Algorithm 2 for aggregator level aggregation can be found in the document.",
    "The content inside Algorithm 2 for aggregator level aggregation includes steps for data preprocessing and feature extraction.",
    "The performance evaluation summary states that the model achieved an accuracy of 90% on the test dataset."
]
```



List of BLEU scores by the OpenAI LLM for multiple user queries



Performance Metrics at varied chunk sizes

CONCLUSION

The integration of Langchain and OpenAI's Language Model (LLM) presents a significant advancement in text processing and question-answering systems. By combining these cutting-edge technologies, the tool offers a comprehensive solution to address various challenges in information retrieval and analysis. Firstly, the tool streamlines the process of text extraction from PDF documents, leveraging Langchain's capabilities for efficient text processing. It automatically splits the text into smaller, manageable chunks and indexes them using FAISS, ensuring fast and efficient retrieval of relevant information from large document collections

FUTURE ENHANCEMENTS

Some potential ideas that can be implemented are :

- Development of a deployment strategy will be conducted that ensures the efficient deployment of the PDF-query tool in the intended environment.

- Robust monitoring and logging mechanisms will be implemented to track system performance metrics in real-time.

- Implementation of comprehensive security measures will be executed to protect sensitive data and ensure compliance with data privacy regulations, utilize encryption techniques to secure data both in transit and at rest.

THANK YOU!