Information Extraction from PDF Documents using Large Language Models

*Dr Maivizhi R Dibyajyoti Ganguly Tarun Negi*

*Assistant Professor, School School of Computing School of Computing*

*of Computing, Srm Institute Srm Institute of Science and Technology Srm Institute of Science and Technology*

*of Science and Technology* *Tamil Nadu, India Tamil Nadu, India*

*Tamil Nadu, India dg7255@srmist.edu.in tn9791@srmist.edu.in*

***Abstract*—**

**Efficiently extracting information from documents/ articles remains a critical challenge in various domains, including academia, industry, and research. 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. In our work, we propose a novel approach leveraging advanced natural language processing techniques for precise information retrieval from PDF documents. Our proposed work leverages the powerful combination of Langchain and OpenAI. This innovative approach is designed to streamline research by allowing users to input articles 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. In this project, we implemented our proposed work, leveraging Langchain, and OpenAI's Language Model (LLM). The experimental results demonstrate the effectiveness of our approach in extracting the information from Portable Document Formats (PDF’s).**

***Keywords— document, natural language processing, pdf,***

***query, prompt, retrieval, response***

I. 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. We aim to simplify this process by offering a user-friendly platform where users can input documents / articles and receive targeted and succinct summaries based on their specified prompts. It extracts relevant facts but also distills the essence of articles, providing users with a comprehensive 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. As digital documents, especially those in the Portable Document Format (PDF), continue to grow at an exponential rate, there is an increasing demand for sophisticated tools that can quickly and precisely search through these documents to find specific information that users are looking for. Our work aims to further the field of question-answering systems by providing a new method designed especially for extracting accurate and exact responses from PDF documents. We hope to provide users with effective access to the wealth of knowledge contained in PDFs by fusing state-of-the-art NLP techniques with careful document preprocessing strategies.

Conventional techniques for obtaining information from PDF documents frequently rely on manual scanning or keyword searches, both of which can be laborious and inaccurate, particularly when working with large or complicated documents. Furthermore, these techniques usually return whole documents or portions of them, meaning that users must sort through potentially irrelevant information in order to locate the exact solutions they need.

In contrast, the proposed model-driven approach aims to streamline this process by enabling users to pose natural language questions (prompts) and receive exact answers directly extracted from the pertinent sections of PDF documents.

The use of cutting-edge natural language processing (NLP) methods, such as deep learning architectures, which have shown amazing aptitude for comprehending and producing text that resembles that of humans, is fundamental to this strategy. Taking advantage of these advances in natural language processing, our approach aims to understand the complex semantics of both questions and documents so that it can correctly identify and extract the most pertinent information.   
  
The paper is structured as follows: Section II reviews related work on information extraction. Section III dives into our proposed work and gives an apt explaination of our

methodology. Section IV focuses on the experimental evaluation of the system and comapres our work with other related works. Finally, we conclude with Section V.

II. RELATED WORK

In recent years, the field of question-answering with PDF document inputs has garnered significant attention due to the ever-expanding volume of digital documents and the growing demand for efficient information retrieval systems. Traditional methods for accessing information within PDF documents, such as keyword searches and manual scanning, often prove inadequate when dealing with extensive or complex documents, leading to inefficiencies and inaccuracies in information retrieval. Recognizing this challenge, researchers have proposed innovative approaches aimed at enhancing the extraction and understanding of information from PDFs. The following review of related work provides insights into the diverse methodologies and advancements made in this domain, ranging from novel approaches for information extraction to the integration of multi-modal techniques for visually-rich document understanding.

Raymond J.Mooney et al. [2] recognized the challenge that while unstructured text contains valuable information, it is difficult to analyze and use effectively without converting it into a structured format. They proposed an approach that combines text mining techniques with information extraction (IE) methods. Information extraction involves identifying specific pieces of information (such as names, dates, events, etc.) from text and structuring them into a usable format. This enables the conversion of unstructured text data into structured formats, making it easier to analyze and use for various applications. However, this approach has problems when a novel extracted entity is represented by similar but not identical strings in different documents and good metrics for evaluating the interestingness of text-mined rules are also needed.

Luis Tari et al. [3] identified the problem in traditional information extraction systems that whenever a new extraction goal emerges or a module is improved, extraction has to be reapplied from scratch to the entire text corpus even though only a small part of the corpus might be affected. In their work, they describe a novel approach for information extraction in which extraction needs are expressed in the form of database queries, which are evaluated and optimized by database systems. Using database queries for information extraction enables generic extraction and minimizes reprocessing of data by performing incremental extraction to identify which part of the data is affected by the change of components or goals. A major drawback of the work is that it does not provide the ability to compute statistics across multiple extraction such as taking redundancy into account for boosting the confidence of an extracted fact.

Muhammad Rio Bastian et al. [4] stated the main problem of rule-based information extraction technique i.e. the extraction rules tend to be specifically designed for specific information or document structure; hence it cannot be directly used in another without some proper modifications. Statistics indicator is a source of information that use tables as a means of data presentation. Statistics indicators also have a relationship concept that must be carefully identified and extracted. Generalization rules attempt to reduce effort in the extraction rule modification process by creating extraction rules in general terms. Combined with ontology, the rules can also extract the relationship between indicators. The output of this information extraction system is a database that keeps not only the data itself but also the relationship concept between indicators. The converted CSV tables in the evaluation phase are not perfect: some columns header are misaligned and the texts are incomplete. When CSV files are perfectly converted, the system achieved 97% accuracy. The inaccuracies are caused by variation of terms that is not covered within ontology.

# Song Liu et al. [5] have proposed that in question answering systems, the greatest hardship is crossing the gap between the questions and corpus to find the appropriate answers. The questions and corpus are in different context, to find the answers the inference whether one candidate sentence fits the user's request is necessary. Pragmatic is concerned with inference about the utterances and user. It is a bridge across the gap of questions and corpus. In their work, they have analyzed the elements of pragmatic and extracted the pragmatic information from the corpus based on the paragraph act. While evaluating, it is observed that the appended pragmatic information significantly improves the performance of the Question Answering System. It should be noted that analyzing paragraph acts and incorporating pragmatic information adds complexity to the information extraction process, requiring sophisticated natural language processing techniques.

# Information Retrieval (IR) system finds the relevant documents from a large dataset according to the user query. Queries submitted by users to search engines might be ambiguous, concise and their meaning may change over time. As a result, understanding the nature of information that is needed behind the queries has become an important research problem. Naw Thiri Wai Khin et al. [1] have proposed the Web Query Classification Algorithm by using NoSQL graph database. This system classifies the web queries into each characteristic and each predefined target categories. In web query classification, the input query is first classified into web search taxonomies (characteristics). Then, domain terms are extracted from the query, and each of them is classified into their relevant categories that are stored in the NoSQL database. However, if the query length is too long and the query contains spelling errors, then this system can’t correctly classify the user query.

# Imran Rasheed et al. [6] have proposed a novel approach in association with the need of constantly upgrading the information retrieval system to meet the challenges posed by the advanced user queries as the search system becoming more sophisticated with time i.e. to augment the query where the automatic query expansion increases the precision in information retrieval even if it can cut down the results for some queries. This approach was tested with the present Urdu data collection obtained via different expansion models such as KL, Bo1 and Bo2. In all the cases addressed, the Bo1 and KL models performed almost similarly whereas the results obtained with Bo2 is not appreciable because lesser number of relevant documents are retrieved due to the term mismatch issues between the original query terms and the candidate expansion terms.

# The proposed research across various information retrieval tasks tackles limitations of current methods. It leverages advanced techniques such as Open Information Extraction (OIE) using an attention-based bidirectional long short-term memory neural network model, which allows for more sophisticated analysis of text data compared to traditional rule-based approaches. Moreover, unlike some of the related work that focuses on specific domains (e.g., job announcements, statistics indicators ), our approach has a broader scope, can extract information from diverse set of documents and, can better capture complex patterns and relationships in unstructured text data, resulting in more accurate extractions. Our proposed approach can inherently grasp the relationships between words without the need of explicit rule creation and ontology.

III. SEMANTIC DOCUMENT RETRIEVAL

This module explains the workflow of our proposed approach. In our implementation, we begin with the extraction of text from the PDF document, followed by splitting the text into smaller, manageable chunks using langchain. Subsequently, language embeddings are downloaded from OpenAI to facilitate deeper semantic understanding of the text. The text segments are then indexed using the Facebook AI Similarity Search framework, involving the creation of a FAISS index and the addition of text segments to this index for efficient retrieval. Upon receiving a user query, the system initiates a search through the indexed text segments to retrieve the relevant documents. The question-answering model is loaded next to process the user query and relevant documents, ultimately producing an answer. Finally, the system outputs the answer, completing the process of querying and retrieving information from the indexed text segments.

**Algorithm specification: question answering**

**Require:** User query, 𝓓 is the set of documents

**Ensure:** 𝓕is the final answer to the user query

**Function** Information Extraction (𝓓)

𝓕 := ∅

**For** each document 𝓓ᵢ in 𝓓

Extract text from 𝓓ᵢ

Split text into smaller segments using langchain

Download language embeddings from OpenAI

Index text segments using FAISS

Receive user query

Search indexed text segments for relevant documents

Load question-answering model

**For** each relevant document 𝓓ᵢ

Process user query and document using question-answering model

Check the relevance and correctness of the extracted information from the document

Obtain 𝓕(i.e. the final answer) after scanning all the releveant documents

**Return** 𝓕

Firstly, the algorithm depicts that each document is processed in the document set. The text from each document is extracted and broken into smaller fragments for better handling. Then, indices for these text segments are created using a technique called FAISS (Facebook AI Similarity Search). This allows the quick and faster search for relevant segments when given a user query. When a user submits a query, the search is initiated through the indexed text segments to find documents that are likely to contain relevant information related to the query. For each relevant document found, a question-answering model is used to process both the user query and the document's content. After extracting information from each relevant document, the relevance and correctness of the extracted information is checked subsequently. It ensures that the information accurately addresses the user's query and is reliable. Finally, after scanning through all relevant documents and extracting information, the answer to the user’s question i.e. the response is obatined.

IV. EXPERIMENTAL RESULTS

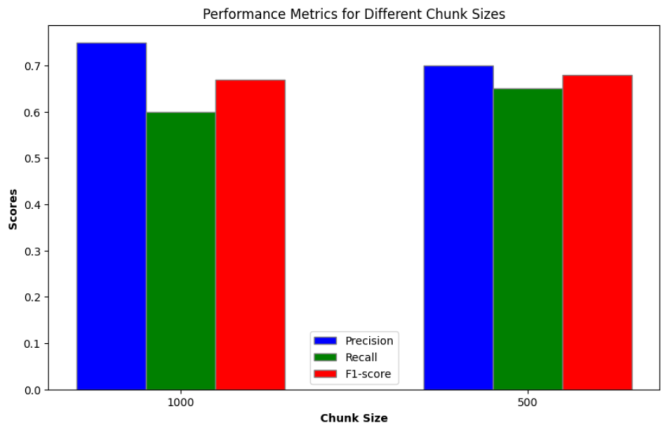
In our research, the entire text of the document is split into chunks. Here, chunk size plays a crucial role in dividing the document into manageable segments. The choice of chunk size is a balance between granularity and efficiency. With a chunk size set to 1000 characters, we ensure that each segment is of a manageable length for processing. Larger chunk sizes may lead to better computational efficiency but also results in the loss of fine-grained details. Conversely, smaller chunk sizes provide more granularity but increases computational overhead. By setting the chunk\_size parameter to 1000, we aim to strike a balance between granularity and efficiency, ensuring effective processing while retaining important contextual information within each segment. Through our proposed approach, we try to ensure that the system delivers accurate and reliable results, particularly in tasks such as information extraction and question answering, where precision is crucial.

We have also used a crucial parameter called chunk\_overlap in our approach. The chunk\_overlap parameter determines the overlap between consecutive chunks of text. Overlapping chunks play a crucial role in maintaining context and coherence during text analysis. In our research, we set the chunk\_overlap parameter to 200, meaning that adjacent chunks overlap by 200 characters. This ensures that important contextual information is not lost at chunk boundaries. Overlapping chunks enable a smoother transition between segments, allowing for a more seamless analysis of the text. By reducing the risk of information loss, overlapping chunks contribute to the overall effectiveness of the text analysis process. Through the integration of Langchain and OpenAI's Language Model, our proposed approach demonstrates effectiveness in extracting textual data, converting it into numerical vectors for semantic analysis, and generating accurate and context-aware responses.

TABLE I. EXPERIMENTAL RESULTS AT VARIED CHUNK SIZES

|  |  |  |  |
| --- | --- | --- | --- |
| **Chunk Size** | **Precision** | **Recall** | **F1-Score** |
| 1000 | 0.75 | 0.60 | 0.67 |
| 500 | 0.70 | 0.65 | 0.68 |

On performing experimental analysis, we obtain interesting insights on the impact of chunk size on the performance of the system. 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 but at the expense of potentially missing out on relevant information. On the other hand, smaller chunk sizes offer a better balance between precision and recall, resulting in a slightly higher F1-score. These findings highlight the importance of carefully tuning the chunk size parameter to optimize system performance based on specific use cases and requirements.

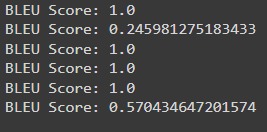


**Fig. 1. Evaluation metrics at varied chunk sizes**

In our proposed work, we emphasize precision because it reflects the accuracy of the system in providing correct answers among all the answers it provides. Precision is particularly crucial when dealing with tasks such as information extraction or question answering, where the goal is to ensure that the provided answers are as accurate as possible. Choosing a chunk size of 1000 over 500 is based on the balance between precision and other metrics such as recall and F1-score. While a chunk size of 500 showed slightly better recall and F1-score, the higher precision achieved with a chunk size of 1000 indicates that the system is more accurate in providing correct answers. In scenarios where accuracy and precision are paramount, such as information retrieval from PDF documents or generating summaries, prioritizing higher precision can be beneficial. Therefore, in our proposed work, we choose a chunk size of 1000 to maximize precision and ensure that the system delivers accurate and reliable results, aligning with the objectives of our research.

Moreover, in our work, we have also implemented another evaluation metric which is far more suitable comapred to precision, recall etc. for text genreation tasks where the entire response to the user query is generated. This performance metric is known as the BLEU (Bilingual Evaluation Understudy) score. This score serves as a valuable metric for evaluating the quality of machine-generated text. In our research, we utilize the BLEU score to assess the quality of answers generated by the OpenAI Language Model. By comparing machine-generated answers with human translation references (generated responses and ground-truth responses), the BLEU score provides a quantitative measure of grammatical correctness and factual accuracy. Higher BLEU scores indicate a closer alignment with human references, reflecting the overall effectiveness of the generated responses. Through the use of the BLEU score, we gain valuable insights into the performance of the OpenAI Language Model in generating accurate and contextually relevant answers. 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 experimental results demonstrate the system's ability to accurately comprehend and analyze documents, delivering relevant information based on user-defined prompts. Through qualitative evaluation, we assessed the system's performance in providing precise summaries tailored to users' requirements. When responding to user queries, the tool utilizes a question-answering model loaded from OpenAI to generate answers based on the retrieved documents. However, the tool's responses are limited to the information available in the document.



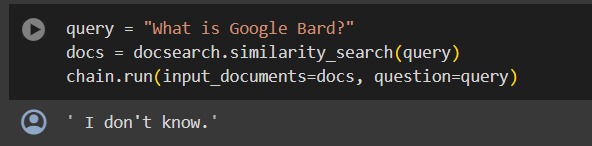
**Fig. 2. BLEU scores coressponding to various user queries and generated responses**

The BLEU score that we have used in our approach is better than the evaluation metrics used in other realted works in certain aspects. This performance metric is more relevant since our work significantly reformulates retrieved information to create the answer, with grammatical correctness alongside factual accuracy. For open ended questions where the answer can be phrased in multiple ways, BLEU provides insight into how well the answer captures the overall meaning of the relevant retrieved information. Following this, we have also shown few of the user prompts and generated answers below.



**Fig. 3. Queries and corresponding responses to data contained in documents**

For example: When asked " What is Google Bard? ", the system responds with "I don't know." This indicates that the document does not contain information about Google Bard, and therefore, it cannot provide a meaningful response, which can be seen below.

**

**Fig. 4. Queries and corresponding responses to data not contained in documents**

Our proposed approach leverages Langchain and OpenAI's Language Model to effectively extract and analyze information from PDF documents. Experimental results demonstrate the tool's effectiveness in extracting textual data, converting it into numerical vectors for semantic analysis, and indexing processed text segments for efficient retrieval. Through qualitative evaluation, we assessed the tool's performance in providing precise summaries tailored to users' requirements, leveraging state-of-the-art language models for accurate and context-aware responses.

The integration of Langchain and OpenAI's Language Model enhances the tool's capabilities, enabling fast and accurate retrieval of relevant information and generating accurate answers based on retrieved documents. Compared to traditional models, this approach benefits from the advanced capabilities of deep learning models, contributing to the development of a robust and efficient information retrieval system.

V. CONCLUSION

According to the experimental results, our approach excels in handling PDF documents, which are a common format for storing and sharing text-based information. Furthermore, our use of the BLEU score as a metric for evaluating the quality of machine-generated text provides valuable insights into the performance of the OpenAI Language Model. The BLEU score allows us to assess the grammatical correctness and factual accuracy of the generated answers, ensuring that the responses align closely with human references. The primary limitation of our work is its ability to processing only PDF documents. This restriction may hinder its applicability in scenarios where information is stored in other formats, such as Word documents, HTML files, or structured databases. Users may need to convert their data to PDF format before utilizing the system, leading to inconvenience and potential data loss. These limitations will be looked into and resolved in future enhancements.

REFERENCES

x

|  |  |
| --- | --- |
| [1] | Nyo Nyo Yee Naw Thiri Wai Khin, "Query Classification based Information Retrieval System," in *ICIIBMS 2018, Track 2: Artificial Intelligent, Robotics, and Human-Computer Interaction*, Bangkok, Thailand, 2018. |
| [2] | Un Yong Nahm Raymond J. Mooney, "Text Mining with Information Extraction," in *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium*. |
| [3] | Phan Huy Tu, Jo¨ rg Hakenberg, Yi Chen, Tran Cao Son, Graciela Gonzalez, Chitta Baral Luis Tari, "Incremental Information Extraction Using Relational Databases," in *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1*, 2012. |
| [4] | Ayu Purwarianti Muhammad Rio Bastian, "Information Extraction in Statistics Indicator Tables using Rule Generalizations and Ontology," in *2016 International Conference on Information Technology Systems and Innovation*, 2016. |
| [5] | Fuji Ren Song Liu, "PARAGRAPH ACT BASED PRAGMATIC," in *Proceedings of IEEE CCIS2011*. |
| [6] | Haider Banka Imran Rasheed, "Query Expansion in Information Retrieval for Urdu Language," in *2018 Fourth International Conference on Information Retrieval and Knowledge Management*, 2018. |
| [7] | Sachin S. Patil Rasika P. Saste, "Extraction of Incremental Information Using Query Evaluator," in *2014 IEEE*. |
| [8] | Christina Feilmayr, "Text Mining-Supported Information Extraction," in *2011 22nd International Workshop on Database and Expert Systems Applications*, 2011. |
| [9] | Morteza Fayazi, Zineb Benameur-El, Shuyan Yu, Ronald Dreslinski Serafina Kamp, "Open Information Extraction: A Review of Baseline Techniques, Approaches, and Applications," , 2023. |
| [10] | Z.Chen, X. Lu W. Yue, "Using Query Expansion and Classification for Information Retrieval," , Proceedings of the First International Conference on Semantics, Knowledge, and Grid (SKG), IEEE, 2006. |
| [11] | Y. F. Hassan S. M. Fathalla, "A Hybrid Method for User Query Reformation and Classification," in *IEEE, pp. 132-138, 2012.* |
| [12] | X. Wang C. Xia, "Graph Based Web Query Classification," , 12th Web Information System and Application Conference, IEEE, pp. 241-. |
| [13] | T. Baldwin S. M. Kim, "An Unsupervised Approach to Domain-Specific Term Extraction," in *University of Melbourne, Australia, 2011.* |
| [14] | M. M. T. Naing, "Ontology-Based Web Query Classification for Research Paper Searching," in *International Journal of Innovations in Engineering and Technology (IJIET) , vol. 2, no. 1, 2013*. |

x