ESG Accountability Made Easy: DocQA at Your Service

Lokesh Mishra[1], Cesar Berrospi[1], Kasper Dinkla[1], Diego Antognini[1], Francesco Fusco[1], Benedikt Bothur[2], Maksym Lysak[1], Nikolaos Livathinos[1], Ahmed Nassar[1], Panagiotis Vagenas[1], Lucas Morin[1,3], Christoph Auer[1], Michele Dolfi[1], Peter Staar[1]

[1]IBM Research, Rüschlikon, Switzerland
[2]IBM Technology, Zürich, Switzerland
[3]ETH Zürich, Zürich, Switzerland
{mis, ceb, dkl, ffu, mly, nli, ahn, pva, lum, cau, dol, taa}@zurich.ibm.com, {Benedikt.Bothur, Diego.Antognini}@ibm.com

Abstract

We present Deep Search DocQA. This application enables information extraction from documents via a question-answering conversational assistant. The system integrates several technologies from different AI disciplines consisting of document conversion to machine-readable format (via computer vision), finding relevant data (via natural language processing), and formulating an eloquent response (via large language models). Users can explore over 10,000 Environmental, Social, and Governance (ESG) disclosure reports from over 2000 corporations. The Deep Search platform can be accessed at: https://ds4sd.github.io.

Introduction

The global impact of climate change has galvanized organizations to announce key information about their environmental footprint (carbon emissions, energy usage, waste emission and management, etc.). Integrating sustainability information into the company reporting cycle is one of the targets of the UN 2030 Agenda for Sustainable Development and institutions like Principles for Responsible Investing (a UN-supported network of investors) encourage investors to incorporate this information into their investment decisions . Companies are thus increasingly disclosing environmental, social, and governance (ESG) data in their ESG reports, typically as PDF files.

Unlike financial data, regulators such as the U.S. SEC do not require public companies to file ESG data with specific forms. There have been massive efforts from several organizations to standardize these reports. However, major challenges continue to persist, including complex regulations, rapidly evolving reporting frameworks, verifying ESG compliance, among others. These matters become more complicated when we realize that most of the ESG reporting is done in non machine-readable formats. Unlocking this vast amount of data in an easily consumable manner would greatly help researchers, policy-makers, lawyers, and corporations by extracting information and gaining insights.

To this end, we have developed Deep Search DocQA. The application offers users to perform document *question answering* (QA), i.e., users can extract information from any ESG report in our library via our QA conversational assistant. Our assistant generates answers and also presents the information (paragraph or table), in the ESG report, from which it has generated the response.

Report	Question	Answer
IBM 2022	How many hours were spent on employee learning in 2021?	22.5 million hours
IBM 2022	What was the rate of fatalities in 2021?	The rate of fatalities in 2021 was 0.0016.
IBM 2022	How many full audits were conducted in 2022 in India?	2
Starbucks 2022	What is the percentage of women in the Board of Directors?	25%
Starbucks 2022	What was the total energy consumption in 2021?	The total energy consumption in 2021 was 2,491,543 MWh.
Starbucks 2022	How much packaging material was made from renewable materials?	31% of packaging materials were made from recycled or renewable materials in FY22.

Table 1: Example question answers from the ESG reports of IBM and Starbucks using Deep Search DocQA system.

Related Work

The DocQA integrates multiple AI technologies, namely:

Document Conversion: Converting unstructured documents, such as PDF files, into a machine-readable format is a challenging task in AI. Early strategies for document conversion were based on geometric layout analysis (Cattoni et al. 2000; Breuel 2002). Thanks to the availability of large annotated datasets (PubLayNet (Zhong et al. 2019), DocBank (Li et al. 2020), DocLayNet (Pfitzmann et al. 2022; Auer et al. 2023), deep learning-based methods are routinely used. Modern approaches for recovering the structure of a document can be broadly divided into two cate-

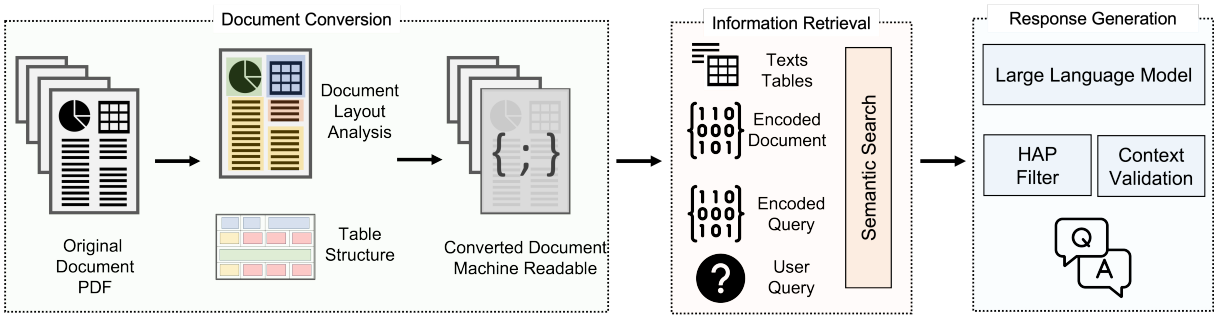

Figure 1: System architecture: Simplified sketch of document question-answering pipeline.

gories: *image-based* or *PDF representation-based*. Image-based methods usually employ Transformer or CNN architectures on the images of pages (Zhang et al. 2023; Li et al. 2022; Huang et al. 2022). On the other hand, deep learning-based language processing methods are applied on the native PDF content (generated by a single PDF printing command) (Auer et al. 2022; Livathinos et al. 2021; Staar et al. 2018).

Application of NLP to ESG: ESG reports contain large amount of useful data in textual and tabular format. There have been some attempts to use NLP on this data. Luccioni, Baylor, and Duchene (2020) developed ClimateQA, a model trained to classify whether a sentence from an ESG report answers regulatory questions. In addition, there are several works which aim to mine information from ESG reports for financial predictions (Guo et al. 2020; Goel et al. 2020). Nevertheless, to the best of our knowledge, no QA system which can extract data directly from an ESG report PDF has been reported in the literature.

LLM & RAG: Due to the increasing scale of training data and model size, large language models (LLMs) demonstrate surprising emergent properties (Wei et al. 2022). For example, the behaviour of the GPT-3 model, with 175 billion parameters, could be modified with in-context learning (Brown et al. 2020; Bommasani et al. 2022; Raffel et al. 2020). Such LLMs are adaptable to a variety of downstream tasks via prompting and can be fine-tuned to better perform in a specific ESG domain (Webersinke et al. 2022). The Retrieval Augmented Generation (RAG) approach aims at improving the performance of these models on knowledge intensive tasks (Lewis et al. 2020). In this approach, the capabilities of natural language generation are combined with a knowledge index, from which relevant documents are retrieved.

System Architecture

In this section, we describe the AI technologies which are integrated into our document question-answering application. The architecture is described in Fig. 1. The pipeline works end-to-end from PDF documents to question-answering using LLMs. It consists of three components described below.

Document Conversion: The document conversion system is designed in an asynchronous task-based queue-worker architecture. The user-facing API accepts documents in PDF format (both programmatically created and scanned). The client receives a task identifier, while an orchestrator enqueues several ML tasks to ephemeral workers. After splitting the document into pages, we: 1) depending on the nature of the PDF, we employ either PDF parsing or OCR, 2) analyze layout and segment it (Auer et al. 2022; Livathinos et al. 2021; Staar et al. 2018) and 3) extract table structures (Lysak et al. 2023; Nassar et al. 2022). Finally, the data from multiple pages is assembled together, preserving the reading order, in a machine-readable format.

Information Retrieval: Using an encoder model, vector embeddings for the data in a document are computed and stored in a vector database. For text this is relatively straightforward, for tables the triplet of (cell content, column header, row header) is expressed as a sentence which gets encoded. The sentence expression is: `string(column header) + string(row header) = string(cell content)`[1]. We perform a k-nearest neighbour search to identify the top-k relevant passages for a user query. For sentence encoding, we use several encoding models from the Sentence Transformer library (Reimers and Gurevych 2019).

Response Generation: We employ a suite of LLMs like LLAMA 2 (Touvron et al. 2023), Flan-UL2 (Tay et al. 2023), or T5 (Raffel et al. 2020) for generating a response to the user query. The user query and relevant context (identified by the previous model) are packaged together in a prompt for the LLM. The response of the model is checked against hate speech, abuse, and profanity. Finally the response is grounded in the context and inspected for hallucinations. If all tests are passed, the response is presented to the user via a virtual assistant. Table 1 shows some examples of questions and the generated answers by the system.

Conclusions

In this paper, we presented our DocQA application targeting ESG reports. The DocQA system can be useful for anyone, from policy-makers to students, trying to find information from a large document. Our future work is focused on enabling querying on multiple documents at once to extract aggregated insights for questions like "How have the Scope 1 emissions evolved over the last decade?". In addition, we will expand this service to other types of documents like scientific papers, financial reports, and patents.

[1]Here, string() returns the string representation of an object.

References

Auer, C.; Dolfi, M.; Carvalho, A.; Ramis, C. B.; and Staar, P. W. J. 2022. Delivering Document Conversion as a Cloud Service with High Throughput and Responsiveness. In *2022 IEEE 15th International Conference on Cloud Computing (CLOUD)*. IEEE.

Auer, C.; et al. 2023. ICDAR 2023 Competition on Robust Layout Segmentation in Corporate Documents. In *Lecture Notes in Computer Science*, 471–482. Springer Nature Switzerland.

Bommasani, R.; et al. 2022. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.

Breuel, T. M. 2002. Two Geometric Algorithms for Layout Analysis. In Lopresti, D.; Hu, J.; and Kashi, R., eds., *Document Analysis Systems V*, 188–199. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-45869-2.

Brown, T. B.; et al. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Cattoni, R.; Coianiz, T.; Messelodi, S.; and Modena, C. 2000. Geometric Layout Analysis Techniques for Document Image Understanding: a Review. Technical Report TR9703-09, ITC-irst, Via Sommarive 18, I-38050 Povo, Trento, Italy.

Goel, T.; Jain, P.; Verma, I.; Dey, L.; and Paliwal, S. 2020. Mining company sustainability reports to aid financial decision-making. In *The AAAI-20 Workshop on Knowledge Discovery from Unstructured Data in Financial Services*.

Guo, T.; et al. 2020. ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction. arXiv:2005.02527.

Huang, Y.; et al. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *Proceedings of the 30th ACM International Conference on Multimedia*.

Lewis, P.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Li, J.; et al. 2022. DiT: Self-supervised Pre-training for Document Image Transformer. *Proceedings of the 30th ACM International Conference on Multimedia*.

Li, M.; Xu, Y.; Cui, L.; Huang, S.; Wei, F.; Li, Z.; and Zhou, M. 2020. DocBank: A Benchmark Dataset for Document Layout Analysis. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 949–960. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Livathinos, N.; Berrospi, C.; Lysak, M.; Kuropiatnyk, V.; Nassar, A.; Carvalho, A.; Dolfi, M.; Auer, C.; Dinkla, K.; and Staar, P. 2021. Robust PDF Document Conversion using Recurrent Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17): 15137–15145. Number: 17.

Luccioni, S.; Baylor, E.; and Duchene, N. 2020. Analyzing Sustainability Reports Using Natural Language Processing. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Lysak, M.; Nassar, A.; Livathinos, N.; Auer, C.; and Staar, P. 2023. Optimized Table Tokenization for Table Structure Recognition. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part II*, 37–50. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-41678-1.

Nassar, A.; Livathinos, N.; Lysak, M.; and Staar, P. 2022. TableFormer: Table Structure Understanding with Transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4604–4613. Los Alamitos, CA, USA: IEEE Computer Society.

Pfitzmann, B.; et al. 2022. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 3743–3751. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9385-0.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Staar, P. W. J.; et al. 2018. Corpus Conversion Service. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

Tay, Y.; et al. 2023. UL2: Unifying Language Learning Paradigms. In *The Eleventh International Conference on Learning Representations*.

Touvron, H.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Webersinke, N.; Kraus, M.; Bingler, J.; and Leippold, M. 2022. ClimateBERT: A Pretrained Language Model for Climate-Related Text. In *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.

Wei, J.; et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. Survey Certification.

Zhang, M.; et al. 2023. WeLayout: WeChat Layout Analysis System for the ICDAR 2023 Competition on Robust Layout Segmentation in Corporate Documents. *ArXiv*, abs/2305.06553.

Zhong, X.; et al. 2019. PubLayNet: Largest Dataset Ever for Document Layout Analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1015–1022.