Biotech Batch Record Digitization: Challenges and Role of AI – Part 2

Experimental Setup

The main objective of the experiments was to study and analyze the best possible approach in automating the data extraction process from scanned documents archived in PDF formats, which would also include the handwritten texts.

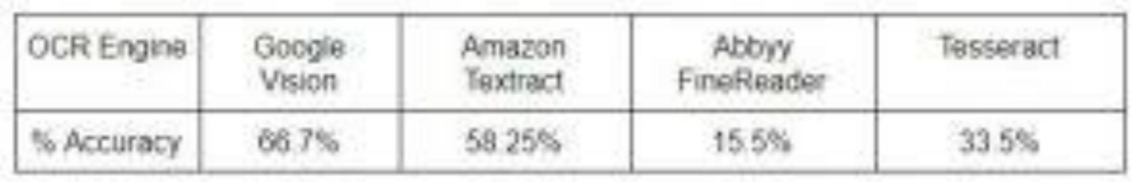
Data: The experiment is performed on a set of sample scanned PDF files. Each PDF file would represent one batch-record. Multiple batch records are obtained from the same CMO and for the same drug, which would make sure that all the PDFs have the same parameters, which can be extracted from them. Also, this ensures that all the PDFs try to closely follow a certain template, which can be thus used in automated software.

OCR Evaluations: The experiment started with the evaluation of the performance of various state-of-the-art commercial and open-source OCR engines, and the implementation of various methods to achieve better performances using image processing techniques. The OCRs are evaluated based on factors like ease of use in the BioTech and Pharmaceutical Industry; and the accuracy of performance. Commercial OCRs included in this experiment include Google Vision APIs, Amazon Textract, and Abbyy FineReader, while Tesseract OCR will be used under the open-source OCR category. Based on the performance achieved from the OCRs and its applicability in the BioTech and Pharmaceuticals industry, the next steps would be decided.

Speech-to-text / Speech Recognition: The Speech-to-text solution is analyzed for its applicability in the industry, and the solution is evaluated based on the performance achieved in terms of accuracy of data extraction and speed of data extraction by performing a time study. For evaluating the achievable speed via speech-to-text based data extraction system, the software is created for the workflow which allows to user to choose the PDF files and load extraction metadata, and auto navigates using an intelligent page finding algorithm, simultaneously allowing the user to record his words for data entry into the database against a parameter, thus automating the data structuring process. The time study is performed on five batch records, where each batch record is split into three PDFs. Each PDF denotes a stage in the drug manufacturing process.

OCR Assessment

The OCRs selected for the experimental study included Google Vision, Amazon Textract, Abbyy FineReader, and Google's open-source Tesseract OCR engines. To evaluate the performance of the OCRs, 20 pages from the batch record PDFs in the form of images, were sent to each OCR after cropping the images to sections containing only handwritten information and the obtained text was compared to the ground truth texts which were recorded manually for these pages. The standard text-to-text comparison algorithms were used, in which two strings are compared with each other based on character-wise similarities. The results from the first evaluation are tabulated below.



OCR Engine	Google Vision	Amazon Textract	Abbyy FineReader	Tesseract
% Accuracy	66.7%	58.25%	15.5%	33.5%

It can be observed that Google Vision's API service provides the best results on the handwritten texts but is capable of giving results with only 66.7% accuracy. In the BioTech and Pharmaceuticals domain, the achieved accuracy is much lower than desired, as the information is very sensitive to safety factors. To improve the accuracy of Google's options, i.e. Vision API and Tesseract, we can perform image enhancement techniques, which would provide better images for the OCR, and hence the OCR output accuracy can be improved.

The image processing techniques involved with document processing were implemented as a pre-processing step. Techniques like image sharpening, contrast and brightness adjustments, noise removal, etc. were implemented, and the same images were again passed on to the OCRs. These image improvement techniques proved beneficial for the Tesseract OCR but reduced the accuracy of Vision API. The accuracy from Google Vision API reduced from 66.7% to 60%, but for Tesseract OCR the accuracy increased from 33.5% to 44.1%. This showed that one cannot rely solely on the image processing techniques, but the improvements in the OCR engine itself were required. As Google's Vision API is a third party solution, one has no control over the training of the system or even with data privacy. Hence, for the next round, only Tesseract OCR was considered. It was trained further on handwritten data. The handwritten sections from the PDFs were split into train and test. The training data was used for further training of Tesseract OCR, while the test data was used for evaluating the OCR performance while training.