

paluno
The Ruhr Institute for Software Technology
Institut für Informatik und Wirtschaftsinformatik
Universität Duisburg-Essen

Master project

**Automated Data Extraction from
Transactional Business Documents**

**Evaluation of the Performance of Commercially Available
Tools on Invoices**

Automatisierte Datenextraktion aus Transaktionsgeschäftsdokumenten

Bewertung der Leistung von kommerziell verfügbaren Werkzeugen auf
Rechnungen

Georgi Georgiev
3046160

Essen, 17.03.2021

Betreuung: Wilhelm Koop (M.Sc.)
Erstgutachten: Prof. Dr. Volker Gruhn

Studiengang: Angewandte Informatik- Systems Engineering (M.Sc.)

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt habe. Ich habe alle Stellen, die ich aus den Quellen wörtlich oder inhaltlich entnommen habe, als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Essen, am 15.03.2021

Zusammenfassung

Handel und Dokumente sind eng miteinander verbunden, da der Handel für verschiedene rechtliche und finanzielle Aufgaben von den Dokumenten abhängt. Eigentlich, macht das schnelle Tempo der modernen Welt und das ständig wachsende Volumen an Dokumenten ihre manuelle Verarbeitung zu einer ziemlich langsamen und mühsamen Aufgabe.

Diese Arbeit beschreibt den automatischen Datenextraktionsprozess im Bezug auf Rechnungen (eine Art Transaktionsdokument, der häufig in Unternehmen verwendet wird), der für Mikro- und Kleinunternehmen relevant ist, sowie eine umfassende Untersuchung der derzeit auf dem Markt verfügbaren kommerziellen Datenextraktionssoftwaretools. Es wird die Aufgabe des Extrahierens von Datenfeldern (z. B. Rechnungsabsender, Rechnungsbetrag, gekaufter Service oder Produkt) aus den Formularen sowie einen direkten Vergleich zwischen den Tools hinsichtlich ihrer Eigenschaften und Leistung beschrieben. Die Transaktionsdokumente können heterogen umrissen sein, ohne das tatsächlich gesetzliche Anforderungen an Form und Struktur gestellt werden, was die Aufgabe der Informationsextraktion aus ihnen herausfordernd macht.

Insbesondere werden die vorhandenen Tools im Hinblick auf die Richtigkeit und Vollständigkeit der extrahierten Daten bewertet. Zu diesem Zweck werden in dieser Arbeit die folgenden Metriken verwendet: precision (z. B. aus den als "Absender" klassifizierten Feldern, wie viele tatsächlich "Absender" -Felder sind), recall (z. B. aus den "Absender" -Feldern, wie viele von ihnen korrekt als "Absender" -Felder klassifiziert sind) sowie das F1 score, der den gewichteten Durchschnitt der beiden vorherigen in einem harmonischen Mittelwert kombiniert.

Um die Bewertung durchzuführen, wurden 50 Rechnungen in englischer Sprache mit bekannten und unbekannten Layouts für das Extraktionssystem ausgewählt. Die Extraktionseffizienz jedes Werkzeugs wurde für jedes der folgenden Dokumentschlüsselfelder bewertet: Absendername, Absenderadresse, Empfängername, Empfängeradresse, Rechnungsnummer und Rechnungsdatum.

Die aus den Rechnungsbelegen extrahierten Schlüssel-Wert-Paare mussten mit den oben genannten Maßnahmen bewertet werden. Die Ergebnisse des Benchmarks liegen in Form einer Tabelle vor und geben Auskunft über folgende Fragen:

- Wie gut die Tools mit der zugehörigen Aufgabe umgehen?
- Wo genau versagen sie?

- Wie sich die Systeme in ihrer Konfiguration unterscheiden - was von dem Benutzer als Eingabe benötigt wird
- Welche Faktoren können die Schlüsseldatenextraktion möglicherweise negativ und positiv beeinflussen?
- Was könnte verbessert werden?

Die allgemeine Schlussfolgerung lautet, dass Parashift mit allen an der Bewertung einbezogenen Tools in nahezu jedem Aspekt die besten Ergebnisse erzielt hat. Insbesondere wurden die höchsten Precision Werte in 4 Feldern und 5 Feldern hinsichtlich des Recall von den vollständigen 7 Feldern erreicht. Der F1-Score war auch der höchste in 5 Feldern und die Fehlerrate war die niedrigste - nur 10%, quasi 37 Fehler in 350 Feldern insgesamt, was fast dreimal niedriger war als beim zweiten Tool Syphit mit 29% und dem dritten Google AI mit 31%.

Es wurden die Ergebnisse und die spezifischen Probleme, die in den kommerziellen Tools aufgetreten sind (z.B. nicht erkannte/falsch extrahierte numerische- oder Textfelder; spärliche Segmentierung zwischen Entitäten) weiter analysiert und Hinweise für mögliche Verbesserungen und zukünftige Arbeiten gegeben.

Nach unserem Kenntnisstand, wurde ein solcher Software-Benchmark in Kombination mit der Dokumentenanalyse in einer einzigen Studie noch nicht erstellt. Es gibt mehrere wissenschaftliche Arbeiten, in denen einige der konkurrierenden Tools verglichen werden. Alle Dokumente stammen jedoch von Unternehmen, bei denen solche Softwaretools vorhanden sind und deren Metriken für den Vergleich unterschiedlich sind.

Abstract

Business and documents are closely related to each other since the business depends on them for diverse legal and financially associated purposes. In fact, the rapid pace of the modern world and the constantly growing volume of documentation makes their manual processing a rather slow and tedious task.

In this work, we describe the automatic data extraction process regarding invoices (a type of transactional document frequently used in business), relevant for micro and small businesses alongside with comprehensive research of the commercial data extraction software tools currently available on the market. We describe the task of extracting data fields (e.g. invoice sender, invoiced amount, bought service or product) from the forms as well as a direct comparison between the tools concerning their characteristics and performance. The transactional documents can be heterogeneous in outline with no actual legal requirement about form and structure, which makes the task of information extraction from them challenging.

Specifically, we have evaluated the existing tools with regards to the correctness and completeness of the extracted data. For that purpose we have used the following metrics: precision (e.g. from the results classified as "sender" fields, how many are actually "sender" fields), recall (e.g. from the "sender" fields, how many of them are correctly classified as "sender" fields) as well as the F1 score, which combines the weighted average of the previous two in a harmonic mean.

To execute our evaluation, we have selected 50 invoices in English language with both familiar and unfamiliar layouts for the extraction system. Every tool's extraction efficiency was assessed for each of the following document key fields- Sender name, Sender address, Receiver name, Receiver address, Invoice number and Invoice date.

The extracted key-value pairs from the invoice documents had to be assessed against the metrics, mentioned above. The results of the benchmark are in a form of a table and they provide answers to the following questions:

- How well the tools deal with the associated task?
- Where exactly do they tend to fail?
- How the systems differ in their configuration – what is required by the user as input?

- What factors could potentially influence the key data extraction- both negatively and positively?
- What could be improved overall?

The general conclusion was that from all of the tools included in the assessment, Parashift achieved the best results in almost every aspect. Moreover, it has accomplished the highest values of precision in 4 fields and the highest values of recall in 5 fields out of the complete 7 fields. The F1 Score was also the highest in 5 of the fields and the error rate was the lowest- only 10% or precisely 37 errors in 350 fields overall, which was almost 3 times lower as an error rate than the second tool- namely Syphit with 29 % and the third tool Google AI with 31%.

We have further analyzed the results and the specific problems that occurred in the commercial tools (such as incorrect/unrecognized extracted numerical or textual fields; sparse segmentation between entities etc.) and gave hints for possible improvements and future work.

To the best of our knowledge, such software benchmark combined with document analytics in a single study hasn't been compiled yet. However, there are multiple papers comparing some of the competitor solutions. All of them, though, are written from the companies, which also possess such software tools and the metrics used for the comparison are different.

List of figures

Figure 1: Types of digital data and their differences in structure	2
Figure 2: Typical processing pipeline	6
Figure 3: Docparser pipeline	7
Figure 4: Cloudscan pipeline	8
Figure 5: Precision and Recall formulas	11
Figure 6: Confusion matrix	11
Figure 7: F1 Score formula	11
Figure 8: Example of key-value fields in an EU invoice	16
Figure 9: Example of Field label/ Field value extraction	20

List of tables

Table 1: Involved Tools	5
Table 2: Tested Components	14
Table 3: Resulting values from the combination of field label and field value	21
Table 4: Example Table for each of the fields	22
Table 5: Precision and Recall	25
Table 6: F1 Score	26
Table 7: Examples of each error type	27
Table 8: Error Type benchmark	27

Index

Eidesstattliche Erklärung	II
Zusammenfassung	III
Abstract.....	V
List of figures	VII
List of tables.....	VIII
Index	IX
1 Introduction	1
1.1 Project Focus	2
2 Involved Tools	4
3 Processing pipelines	6
3.1 Typical processing pipeline.....	6
3.2 Processing Pipeline from Transactional Business Documents	7
4 Metrics.....	10
5 Experiment Setup	12
5.1 Tasks	12
6 Invoice Dataset	13
6.1 Challenges and components under test	13
6.2 Invoice Types.....	14
6.3 Targeted fields	15
6.4 The Dataset	17
7 Comparison and Results	20
7.1 Discussion	23
7.2 Error Analysis.....	26
8 Related Work	28
9 Conclusion	30
References	31

1 Introduction

Companies nowadays have to process a large amount of documents every day – from license agreements, through contracts to insurance policies, etc. The vast majority of enterprise information makes data extraction a tedious and rather challenging task for the modern business.

A study by Concur[1] from 2016 has shown, that only 16% of the companies in the UK have fully automated their invoice processing. The study has revealed that an average of 15 employees are tied up in every incoming single invoice and have to manually operate with it. This might not only lead to fraudulent and decelerated invoicing operations, but also to poor visibility, late payments and additional costs. Therefore around 60 % of the employees agree that their organization's invoice operations systems could be improved. Furthermore, they claim they aren't being equipped to handle the large volume of invoices by hand. Each of the mid/large companies, participating in another study from the German Research Center for Artificial Intelligence (DFKI)[2], declares receiving more than 500 invoices per day. An accountant in the company handles around 126 documents every single day and the processing costs per document are 9 Euros. This survey calculated that the percentage of mistyped or misread invoice processing is between 0 and 20 %, with an average of 5%.

We have conducted small sample research, involving several accounting firms serving more than 70 clients, located in Sofia - the capital of Bulgaria, an EU member and one of the emerging economies in South-East Europe. The employees share, that each and everyone of them processes manually 400-700 invoices per day, depending on the additional work and verification that each document requires. This is a labor-intensive and expensive process. In some of the accounting companies we researched, every single document is double-checked by 2 different experts, but even then the error rate is significant, between 5 and 10 %.

Having this information in mind, there is no wonder that many companies nowadays invest in autonomous classification and data extraction tooling, which can ease and accelerate the process of key data extraction. Many of the business processes depend on the invoices- from the process of finding the right supplier to the process of making payments, as well paying out loans and additional charges [3].

The technology advancements, especially in the field of AI (artificial intelligence) and Machine Learning have the potential to improve and speed up document processing, without overloading the company resources and becoming a burden to the employees.

1.1 Project Focus

The main focus of this study was to research the existent commercial tools, as well as to assess the approaches, proposed by the scientific papers in the field of information extraction and natural language processing. (Information extraction (IE) is the automatic retrieval of structured information from a selected source. Sources can be text documents, databases, websites, images, etc. Usually, IE is part of a larger concept - called natural language processing (NLP). NLP can be defined as the ability of a computer program to understand human language. It is a component of artificial intelligence (AI)) [4], [5].

To define the topic more precisely, we have explored the procedure of automated document processing with the help of modern computer technologies and paradigms as well to estimate the type and the average volume of business-related documents, which a typical company processes daily, more particularly invoice documents. Such billing documents are often semi-structured [6] (Fig.1) and include both key-value pairs like for example sellers, buyer's details and the actual payment terms. They can also include a table with the sold items or provided services [3]. Business-related documents differ in layout as well in structure – some of them follow a particular frame, others are in free form text and don't follow any particular composition nor pattern(e.g. employment contracts, business reports, operating agreements, etc.), which makes them difficult to process by the software, because of their diverse nature. In between those types is the semi-structured documentation like invoices, purchase orders, etc.

The project scope was be concentrated mainly on invoices, considering the companies accounting departments, where manual labour of document processing is at most and automatization can be beneficial.

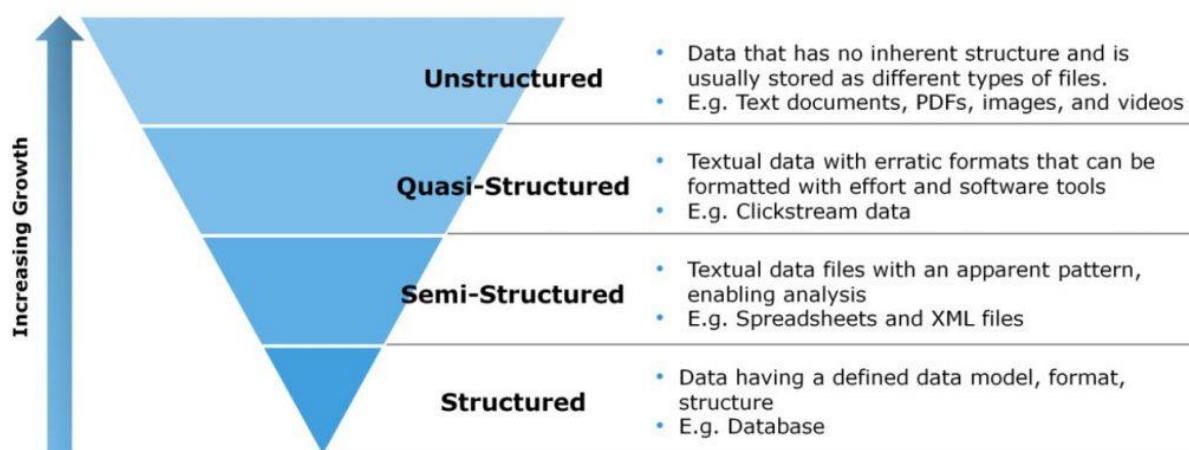


Figure 1: Types of digital data and their differences in structure [7].

The companies that we have focused on as potential clients of information extraction software are small and medium-sized enterprises (SMEs), which number according to a study from 2019 is well over 22 Million in Europe [8].

Despite the growing popularity of the Electronic Data Interchange (EDI), which enables automatic processing and exchange of documents between vendors and clients, most SMEs in Europe are still receiving invoices in paper or PDF format and avoid using it, despite the gained mainstream adaptation of the EDI worldwide [9]. A study [10] from 2017 in Germany, has shown that less than 5% on the issuer side and around 3% on the receiver side use the EDI framework for invoice exchange. Those companies rely heavily on emails and PDFs for this exchange, from which the information has to be extracted later in a structured format, so it can be accessed again if needed.

Another topic of discussion was customer satisfaction, in particular, what could be the important metrics and expectations for a potential customer. Here direct comparison was very helpful and emphasized on the differences between the software tools, as well as the similarities in operation. They differ in working principle, some of them require little to no training to recognize the required fields and some rely heavily on templating and ready-to-use solutions.

We dug deeper into the key data extraction process and examined the individual steps of it. The scientific literature, particularly in the last couple of years, has reported significant advances in the field and reviewing them was very helpful for our research.

We would like to explore the potential benefits of automating the information extraction process in the accounting domain with AI capabilities and specify the type and layout of documents that can be automatically processed with such software. This can help to establish the business value and competence of the tools and possibly give us ideas on what needs to be improved or implemented so that companies will invest and rely on such software in the near future. To do that we need to explore what each product offers as features and functionalities and what are its advantages and disadvantages.

2 Involved Tools

There are numerous information extraction tools available that can help the user perform diverse roles - collect information and draw valuable insights on market research, process employee contracts and job applications, track information from invoices and store them in table form, etc.

Generally, when it comes to the principle of operation functions, we can separate them predominantly into 2 types [11, 12]:

Template-based solutions

The user has to input the document structure, so the machine can later isolate certain parts of the text and process it. This approach is highly accurate if the document matches the coded template [13].

However, these software products have their disadvantages, because:

- Invoice structures change and evolve over time
- The created templates have to be constantly maintained
- A lot of manual annotation is required

Machine Learning-based solutions

The system is trained on different layout formats, therefore the additional work of annotating and setting to the desired template is eliminated. The ML tools used by us are divided into multiple categories:

- **Pre-trained machine learning (ML) solutions:** They are built based on a large dataset of invoices. However, there could be invoices that the software hasn't encountered before.
- **Continuously trained machine learning (ML) solutions:** These solutions are continuously trained on millions of invoices and are most suitable for new and unknown documents.
- **Active learning solutions:** The key idea of such kind of supervised machine learning algorithm is to achieve greater accuracy with fewer labeled training examples when it is allowed to choose the training data from which it learns [14].

We have reviewed and tested the software tools (Table 1) as we discussed their capabilities and how well they complete the assigned tasks. This information is available in chapter 7 "Comparison and Results".

Table 1: Involved Tools

Company	Type of Solution	Website
Google Document AI	ML-based solution (no further details)	https://cloud.google.com/document-ai
Docparser	Template-based solution	https://docparser.com/
Parashift	ML-based solution (no further details)	https://parashift.io/de/
Hypatos	ML-based solution - Continuously trained ML	https://hypatos.ai/de
Rossum	ML-based solution (no further details)	https://rosum.ai/
Sypth	ML-based solution - Active Learning	https://www.sypth.com/

List of the commercial tools, which we have analyzed

3 Processing pipelines

Here, the different approaches to the task of data and information extraction have been explained- the typical pipeline and the pipelines of two commercial products- Docparser and Cloudscan.

3.1 Typical processing pipeline

The standard pipeline for natural language processing and information extraction, described in the scientific literature, comprises the following steps, described below. Depending on the task and type of processed text, it is often part of larger architecture for data extraction, including OCR detection and additional image conversions [15], [16]:

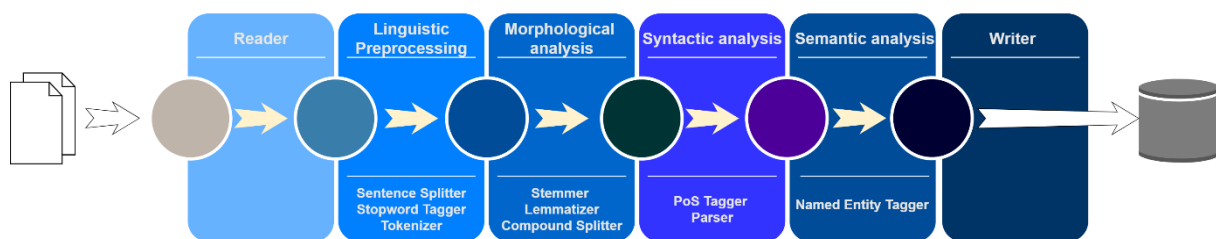


Figure 2: Typical processing pipeline

Initially, a reader component that is capable of processing different formats assigns the documents. After that, follows the preprocessing of the text – assisted by computational linguistics tools such as tokenization, sentence splitting, morphological analysis, etc. In that way, the boundaries of the sentences are identified and each sentence is divided into tokens.

The next step is syntactic analysis, where sentences are segmented into noun groups, verb groups, and particles. The words can be grouped into syntactic phrases with the help of parsing.

Then the semantic analysis is carried out – in that phase is mentions of people, things, locations, events and other pre-specified types of concepts are detected and classified, using the annotations provided from the previous steps. Relationships between the extracted concepts are also being identified.

Before storing the data into the database we have to make sure that we have minimized the duplicate data, thereby minimizing the noise as well.

The final step is improving the knowledge base– this is where the extracted knowledge is stored by a component named writer in the database for further use.

3.2 Processing Pipeline from Transactional Business Documents

The information in this chapter was mainly gathered from two commercial tools. As it can be seen, there is a distinct relation between the standard pipeline definition, explained above and their specific implementations – Docparser and Cloudscan. There are however some differences, since the first approach describes the generic information extraction and the other two are focused more on document-driven data extraction, where some of the procedures from the first can be excess and can cost unnecessary computational power.

Docparser is an end-to-end system for parsing the document's composition- all text structures, figures tables, etc. The systems perform document parsing in these 4 steps [17]:



Figure 3: Docparser pipeline

- 1) **Image Conversion:** The document is converted into an image with a predefined resolution and is resized to a fixed size. These images are further preprocessed and the RGB channels are normalized.
- 2) **Entity Detection:** The image from the previous step is taken and a list of entities is returned. It comprised of the semantic category, coordinates bounding box form and confidence score. This step draws upon on neuronal model for image segmentation and entity detection. This neuronal model determines for each entity its rectangular bounding box, binary segmentation mask responsible for detecting the entity and background pixels and category label of the entity.
- 3) **Relationship classification:** Based on the list of entities from the previous step, the bounding boxes are translated into hierarchical relations with the help of a set of heuristics. Heuristics are differentiated according to whether they generate nesting among entities respectively parent-child relation or are ordered based on the general reading flow.
- 4) **Scalable Weak Supervision:** aims at improving the performance of entity detection and end-to-end parsing. It is built upon an additional dataset that consists of source codes. With the help of that, we can create a mapping between entities in the source code and their renderings. Weak Supervision in Docparser is based on TEX source files that are used to generate document renderings in PDF format. The document structure is stored as a JSON file.

Cloudscan is an invoice analysis system that is capable of learning single global model of invoices and requires zero configuration or upfront annotation. The model is trained with data, extracted from user-provided feedback. The pipeline follows the following order, illustrated below [18]:



Figure 4: Cloudscan pipeline

- 1) **Text Extractor:** PDF document as input. Words and their positions in the PDF are extracted through a standard OCR system. The output is a structured representation of these words and lines in hOCR format (an open standard for representing document layout analysis and OCR results as HTML [19]).
- 2) **N- Grammer:** n-grams (n-gram is a contiguous sequence of n items (phonemes, syllables, letters, words or base pairs) from a given sequence of text or speech. The items can be according to the application. The n-grams typically are collected from a text or speech corpus [20]) of words of the same line are created. Output of this step is a list of N-grams with length up to 4.
- 3) **Feature Calculator:** for every N-gram from the previous step features are calculated. They can be 3 categories: text, numeric or Boolean. Text features can be e.g. raw text of n-gram, all numbers with "0". Numeric features can be e.g. position of the page, width, height, etc. Boolean features can be whether parses as a date or matches known country, city, etc. Output is a feature vector for every N-gram.
- 4) **Classifier:** classifiers all the N-gram feature vectors from the previous step into 32 fields of interest- e.g. invoice number, date, etc. Each N-gram is represented by a vector of 33 probabilities.
- 5) **Post-Processor:** makes a decision on which N-grams will be used for the fields in the output document. First, the N-gram candidates are filtered out, which doesn't fit the syntax of the field after parsing with the related parser. These parsers are based on regular expressions that can find and correct simple OCR errors. For the fields with no semantic connection to other ones, is used Hungarian Algorithm [21]. Output is mapping from fields of interest to chosen N-grams.
- 6) **Document Builder:** invoice is built in Universal Business Language (UBL) (open library of standard electronic XML-based business documents (purchase orders, invoices, etc.) for procurement and transportation. UBL is designed to plug directly into existing business,

legal, auditing, and records management practices [22]) format with values from the N-grams [23].

Information regarding the pipelines and the operating principles of the other tools, participating in the study couldn't be found and wasn't provided by the companies, which own these products. The gathered information about them is compiled and described and Chapter 8" Related Work".

4 Metrics

In summary, there are a fair amount of similarities in the pipeline of these products. A direct comparison has been very helpful and has emphasized on the distinctions between them, as well as the similarities in operation. They differentiate by working principle, some of them require little to no training to recognize the required fields and some rely heavily on templating and ready-to-use solutions.

Based on different characteristics that are important for the customer, a comparison between the approaches was needed. On those grounds, we have considered what would be suitable to be added to information extraction software in the near future.

The aim was to understand how the variety of extraction tools manage various document layouts. Therefore we have performed a set of tests to determine the performance of each of them to see how well they extract particular information from known documents as well from unknown ones. The metrics that we have used are various (precision, recall, f-score, etc.).

By definition, precision (Fig.5) constitutes the proportion of items, returned by the system, which are actually correct. To accomplish high scores, it is needed to carefully select the positives and discard everything, which we aren't completely sure, that is correct. Otherwise, false positives can occur, which affects the precision negatively [24].

On other hand, the recall (Fig.5) specifies how many items from the set that should have been found, are actually found. If high recall is demanded, it is recommended to include everything, even items that we aren't completely confident about. False negatives can occur in either way [24].

The advantage of including them both precision and recall in a benchmark is that one can be more important than the other, depending on the situation and use case. For example, when users search for a piece of particular information in Google, they would like to see what they are looking for, on the very first page of the browser with results relevant only for their search or in other words results with high precision. But when the users search for data on their local hard drives, they would usually search for all of the available data, which meets some criteria, thus striving for higher recall. In our particular case, however, which is more related to information extraction and NLP, the ideal scenario was to achieve high precision and low percentage of false positives, while tolerating a certain amount of recall [25].

Together precision and recall can be similar to each other. To avoid extreme situations and unbalanced results, it is recommended to combine them in a

Single value – F score (Fig.7), or commonly used as F1, hence the coefficient β is typically 1. Values $\beta < 1$ can prioritize precision, while $\beta > 1$ can be lean more towards emphasizing recall [24].

This is how the metrics are calculated:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Figure 5: Precision and Recall formulas [26].

	Negative	Positive
Negative	True Negative	False Positive
Positive	False Negative	True Positive

Figure 6: Confusion matrix [26].

Further, in the chapter “Comparison and Results” we have defined our confusion matrix (Fig.6), from which we have derived True Positive/True Negative and False Positive/False Negative values. We have used them to determine Precision and Recall. Finally, the F1 score was computed (Fig.7).

$$\text{F1} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 7: F1 Score formula [26].

5 Experiment Setup

The tools were tested by a single person from the perspective of an end-user in a period of 3 weeks against the complete invoice dataset on a laptop PC with an internet connection. The user had no prior knowledge about the tested systems, their programmatic principles as well as no accountant expertise.

The assignment for the testing person included finding a sufficient number of invoices, assessing the tools and subsequently protocoling the results in a spreadsheet. These results were used at the end to measure the extraction performance.

The companies, which have manufactured the tested tools, haven't provided any access to source code or internal specifications and no additional services were purchased from the websites of the companies. Only free and trial software versions were used.

5.1 Tasks

The tested data extraction software tools had to properly recognize and extract the relevant fields from invoices, as well to accurately assign them as key-value pairs as a final result.

Examples:

- Important fields like *amount due*, *due date*, *invoice data*, *invoice id*, *purchase order*, *total amount*, *total tax amount*, etc. and their values had to be extracted.
- Systems had to differentiate between vendor or customer names and their distinct branches: *Samsung Electronics*, *Samsung Heavy Industries*, *Samsung Life Insurance*, etc.
- Extraction of information from invoices with a different layout (known & unknown), field position
- Gathering and presenting the final data in comprehensible way

6 Invoice Dataset

6.1 Challenges and components under test

The main challenges in our specific information extraction case arise because it combines both tasks, from the natural language processing (NLP) and computer vision worlds in itself. Unlike them, sometimes the information in an invoice can resemble a form of a table, which makes extracting it even harder. The key to successful implementation lies in the understanding of the 2-dimensional layout of such documents [27].

The tools that we have tested had to deal with a set of “hurdles”, which can occur in real-world practice and are an immutable part of daily document processing (Table 2). First, the images imported by the users can be with inconsistent visual quality- for instance, the document can be damaged before capturing, the photo can be skewed, blurred, at a sub-optimal angle, etc. Second of all, such documentation is characterized by a wide range of visual variability and it doesn’t exist a concise template and layout, nor standardized terminology for the economic phrases, used in the text by the vendors. Additionally, from a security standpoint and considering the sensitive information, which such documents include, it is very demanding to compare the vendors directly as no large and consistent dataset is publicly available on the internet. This hinders the research and advancements in the field and makes it hard to gather valuable information, which later can be used for training or testing machine learning models.

Thus, such complex image analysis problem cannot be solved with a standard OCR solution. Hence, adequate specialized information retrieval tools for document processing are highly demanded, due to their applicability and crucial importance for the business sector. That made the job of the extraction tools more complicated, but it has revealed which one of them reached higher success rates of information extraction when conditions were actually suboptimal.

Table 2: Tested Components

Parameter	Component being Tested	Example
Color, Resolution, Brightness	Text Extractor	e.g. working with colored/black and white photos
Additional obstructions	Text Extractor	e.g. operate despite the presence of handwriting, noise, grain on the image
Layout (tabular vs non-tabular data)	Text Extractor	e.g. separate table from non-table information
Fields, key-value pairs	Entity Classifier, PoS tagger	e.g. correctly identify fields like "invoice number", "date", "due date" and their values, etc.
Identifying/ Differentiating the entity fields	Entity Classifier, N-grammer	e.g. correctly identify "University Duisburg-Essen" as a complete entity

List of the components and the varying parameters under test which we have tested

6.2 Invoice Types

There are different types of invoices that are suitable for various business services. Some of them, applicable especially in the accounting field are [28, 29]:

- Standard invoice
- Credit invoice

- Debit
- Mixed
- Commercial
- Proforma
- Expense
- Final
- Interim

The dataset included documents from multiple international invoice systems: EU, US, UK and Australia. The differences between them appear to be minor but can turn out to be actually detrimental. For example, The EU members and the UK require the charging of VAT (value-added tax) in the sale of products and services, including digital products or any other goods sold online. However, the US system uses sales tax imposed by the government on the sale of goods and services. Nonetheless, the tools which we have chosen in our study, are intended to comply with the international standards for invoicing and can process different types of documents.

6.3 Targeted fields

Regardless of the type of document, some fields are always mandatory like in the case of the EU invoices [30]. In general, for a VAT invoice to be valid, it needs to fulfill the standards set out by the European Commission for Taxation and Customs [31]:

- **Date of issue**
- **Unique sequential number identifying the invoice**
- **Customer's VAT identification number**
- **Supplier's full name & address**
- **Customer's full name & address**
- **Description of quantity & type of goods supplied or type & extent of services rendered**
- **Date of transaction or payment (if different from invoice date)**
- **VAT rate applied**
- **VAT amount payable**
- **Breakdown of VAT amount payable by VAT rate or exemption**
- **Unit price of goods or services – exclusive of tax, discounts**

VAT INVOICE NO. 4056/2016		Invoice number							
Date of issue: 2016-02-29		Date of issue							
Supplier Egger Gmbh Leipziger Strasse 14 Minden Holzhausen Germany DE123456789 32425		Customer Tessier 60, rue Gustave Eiffel 91130 RIS-ORANGIS France FR987654321							
Date of payment: 2016-03-01		Date of payment							
No	Service Description	Quantity	Exchange rate	Unit net price	Net amount	VAT rate	VAT amount	Gross amount	
1	Silver Pen	1	1	15	15	23%	3.45	18.45	
2	Photo Album 100 photos	2	1	150	300	8%	24	324	
3	Photo Album 200 photos	1	1	221	221	10%	22.1	243.1	
Description of quantity of goods supplied					TOTAL	536		49.55	585.55
Breakdown of VAT by rate applied						8%	24		
						10%	22.1		
						23%	3.45		
Payment details: wire transfer, account number DE12500105170648489890									
Payment details									

Figure 8: Example of key-value fields in an EU invoice [32].

Hence, the benchmark was be predominantly concentrated on extracting specific fields and their values from the document, in particular:

- Sender name
- Sender address
- Receiver name
- Receiver address
- Invoice number
- Invoice date
- Total Amount

The list mentioned above is common for all of the documents in our test dataset and is universal for all of the invoice standards around the world.

6.4 The Dataset

The test dataset consisted of 50 invoices in JPG and PNG format from different providers and had diverse layouts and fonts, some of the files contained graphics as well. They were gathered from the RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset which consists of 400,000 grayscale images in 16 classes, including invoices [33] and this Github dataset that contained multiple colored invoice documents [34].

To replicate a real-world scenario and the individual conditions, where pictures can be taken with smartphones or scanned with a scanner, the photos in our dataset were:

- 40 % black and white and 60 % colored documents;

UNIVERSITY OF CALIFORNIA, DAVIS
Accounting Office
Extramural Accounting
Contractors Invoice

INVOICE TO
Dr. Bruce D. Davies
Philip Morris U.S.A.
4201 Commerce Road
Richmond, VA 23334

Invoice Number: 59280-1
Date: October 22, 2001
Amount: \$11,522.10

Questions regarding this invoice should be directed to Evelyn Montoya @ (530) 757-6527
Federal Employee ID # 94-035490V
Contract/Grant/Agreement/Purchase Order # 43314
Period Billed From: 07/01/01 To: 09/30/01
Project Title: Brain Microvessel Culture-Chloride Cotransport in Cerebral Ischemia
PI/Director: Martha E. O'Donnell
Department: Med Human Physiology

Description of Services	Amount
Salaries & Wages	\$6,542.10
Employee Benefits	\$1,904.10
Supplies & Expenses	\$204.13
Travel	\$0.00
Indirect Cost @48%	\$4,162.01
Total Amount	\$12,802.34
Less 10%	(\$1,280.23)
Total Amount Due	\$11,522.10

Please Return Invoice Copy with Check **PAY THIS AMOUNT>>>>** \$11,522.10

Remarks: Outstanding Invoice:

Certified by: Kathleen Hoss, Extramural Funds Manager

Mobile Checks payable only Mail To:
The Regents of the University of California
Cashier's Office
One Shields Avenue
Davis, California 95616

I hereby certify that all expenditures reported (or payments requested) are for appropriate purposes and in accordance with the agreements set forth in the application and award documents.

2085534882

INVOICE

647-444-1234
your@email.com
yourwebsite.com

1 Your Address
City, State, Country
ZIP CODE

Billed To: Client Name 000000
1 Client Address
City, State, Country
ZIP CODE

Invoice Number: 000000
Date Of Issue: 10/07/14


Invoice Total: **\$4520.00**

Description	Unit Cost	Qty / Hr Rate	Amount
Your item Name Item description goes here	\$1000	1	1000
Your item Name Item description goes here	\$1000	1	1000
Your item Name Item description goes here	\$1000	1	1000
Your item Name Item description goes here	\$1000	1	1000

Subtotal \$4000.00
Tax \$520.00

- with resolution ranging from 2500x 2500 pixels to 600x 400;

Sales Receipt



Boost Solutions
We make your Sharepoint life easier

***Make your SharePoint
life Easier.***

Date 8/13/2013

WISHWILL International Limited
 Haidian District, Beijing, P.R. China
 100001
 Phone: +86-10-128041161
 Fax: +86-10-82601161
 sales@boostsolutions.com

SOLD Bill Billy
TO AA88C
 410 South Johnstone
 74003
 Phone: +1-2152322
 Customer ID: CJ-01286-V125K7

Payment Method
Paypal

Check No.
5582248

Qty	Item #	Description	Unit Price	Discount	Line total	
1	ComposeDoc		\$1,559.00	\$313.80	\$1,559.00	
1	MergeDoc		\$4,299.00	\$859.80	\$4,299.00	
1	TransferDoc		\$699.00	\$139.80	\$699.00	
Total Discount				591.340		
					Subtotal	\$4,567.00
					Sales Tax	\$0.00
					Total	\$3,653.60

Thank you for your business!


VERITAS		10101, 10102, 10103, 10104, 10105 10106, 10107, 10108, 10109, 10110 10111, 10112, 10113, 10114, 10115 10116, 10117, 10118, 10119, 10120 10121, 10122, 10123, 10124, 10125 10126, 10127, 10128, 10129, 10130 10131, 10132, 10133, 10134, 10135 10136, 10137, 10138, 10139, 10140 10141, 10142, 10143, 10144, 10145 10146, 10147, 10148, 10149, 10150 10151, 10152, 10153, 10154, 10155 10156, 10157, 10158, 10159, 10160 10161, 10162, 10163, 10164, 10165 10166, 10167, 10168, 10169, 10170 10171, 10172, 10173, 10174, 10175 10176, 10177, 10178, 10179, 10180 10181, 10182, 10183, 10184, 10185 10186, 10187, 10188, 10189, 10190 10191, 10192, 10193, 10194, 10195 10196, 10197, 10198, 10199, 10200 10201, 10202, 10203, 10204, 10205 10206, 10207, 10208, 10209, 10210 10211, 10212, 10213, 10214, 10215 10216, 10217, 10218, 10219, 10220 10221, 10222, 10223, 10224, 10225 10226, 10227, 10228, 10229, 10230 10231, 10232, 10233, 10234, 10235 10236, 10237, 10238, 10239, 10240 10241, 10242, 10243, 10244, 10245 10246, 10247, 10248, 10249, 10250 10251, 10252, 10253, 10254, 10255 10256, 10257, 10258, 10259, 10260 10261, 10262, 10263, 10264, 10265 10266, 10267, 10268, 10269, 10270 10271, 10272, 10273, 10274, 10275 10276, 10277, 10278, 10279, 10280 10281, 10282, 10283, 10284, 10285 10286, 10287, 10288, 10289, 10290 10291, 10292, 10293, 10294, 10295 10296, 10297, 10298, 10299, 10300 10301, 10302, 10303, 10304, 10305 10306, 10307, 10308, 10309, 10310 10311, 10312, 10313, 10314, 10315 10316, 10317, 10318, 10319, 10320 10321, 10322, 10323, 10324, 10325 10326, 10327, 10328, 10329, 10330 10331, 10332, 10333, 10334, 10335 10336, 10337, 10338, 10339, 10340 10341, 10342, 10343, 10344, 10345 10346, 10347, 10348, 10349, 10350 10351, 10352, 10353, 10354, 10355 10356, 10357, 10358, 10359, 10360 10361, 10362, 10363, 10364, 10365 10366, 10367, 10368, 10369, 10370 10371, 10372, 10373, 10374, 10375 10376, 10377, 10378, 10379, 10380 10381, 10382, 10383, 10384, 10385 10386, 10387, 10388, 10389, 10390 10391, 10392, 10393, 10394, 10395 10396, 10397, 10398, 10399, 10400 10401, 10402, 10403, 10404, 10405 10406, 10407, 10408, 10409, 10410 10411, 10412, 10413, 10414, 10415 10416, 10417, 10418, 10419, 10420 10421, 10422, 10423, 10424, 10425 10426, 10427, 10428, 10429, 10430 10431, 10432, 10433, 10434, 10435 10436, 10437, 10438, 10439, 10440 10441, 10442, 10443, 10444, 10445 10446, 10447, 10448, 10449, 10450 10451, 10452, 10453, 10454, 10455 10456, 10457, 10458, 10459, 10460 10461, 10462, 10463, 10464, 10465 10466, 10467, 10468, 10469, 10470 10471, 10472, 10473, 10474, 10475 10476, 10477, 10478, 10479, 10480 10481, 10482, 10483, 10484, 10485 10486, 10487, 10488, 10489, 10490 10491, 10492, 10493, 10494, 10495 10496, 10497, 10498, 10499, 10500 10501, 10502, 10503, 10504, 10505 10506, 10507, 10508, 10509, 10510 10511, 10512, 10513, 10514, 10515 10516, 10517, 10518, 10519, 10520 10521, 10522, 10523, 10524, 10525 10526, 10527, 10528, 10529, 10530 10531, 10532, 10533, 10534, 10535 10536, 10537, 10538, 10539, 10540 10541, 10542, 10543, 10544, 10545 10546, 10547, 10548, 10549, 10550 10551, 10552, 10553, 10554, 10555 10556, 10557, 10558, 10559, 10560 10561, 10562, 10563, 10564, 10565 10566, 10567, 10568, 10569, 10570 10571, 10572, 10573, 10574, 10575 10576, 10577, 10578, 10579, 10580 10581, 10582, 10583, 10584, 10585 10586, 10587, 10588, 10589, 10590 10591, 10592, 10593, 10594, 10595 10596, 10597, 10598, 10599, 10600 10601, 10602, 10603, 10604, 10605 10606, 10607, 10608, 10609, 10610 10611, 10612, 10613, 10614, 10615 10616, 10617, 10618, 10619, 10620 10621, 10622, 10623, 10624, 10625 10626, 10627, 10628, 10629, 10630 10631, 106	
---------	--	--	--

- with varying brightness, contrast and overall image quality;

		JUL 13 1983		RECEIVED FEB0655																																																																
733 THIRD AVENUE, NEW YORK, N.Y. 10017 (212) 480-1200																																																																				
<div style="display: flex; justify-content: space-between;"> <div style="width: 20%; background-color: black; height: 40px;"></div> <div style="width: 40%; background-color: black; height: 40px;"></div> <div style="width: 40%;"></div> </div>																																																																				
<div style="display: flex; justify-content: space-between;"> <div> <input type="checkbox"/> LOEWIS / LOWILLARD ONE PARK AVE. NEW YORK, N.Y. </div> <div> 17TH FL. 10016 </div> </div>																																																																				
<input type="checkbox"/> FOR ADVERTISING IN: PURCHASE IN: STAR-PRESS			<input type="checkbox"/> MONTH OF: JUN 1988																																																																	
<div style="display: flex; justify-content: space-between;"> <div> <input type="checkbox"/> PRODUCT/SERVICE CODE: <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>DATE</th> <th>TOTAL X DFT</th> <th>DATE</th> <th>TOTAL X DFT</th> <th>DATE</th> <th>TOTAL X DFT</th> </tr> </thead> <tbody> <tr> <td>7/10/88</td> <td>4 X 15.75</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>7/11/88</td> <td>6.00-00</td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table> </div> <div> <input type="checkbox"/> ADVERTISER/PRODUCT <input type="checkbox"/> CHARGES <input type="checkbox"/> DESCRIPTION LOEWILLARD 211 HARLEY DAVIDSON CIGARETTES </div> <div> <input type="checkbox"/> GROSS RATE 22.75/500 </div> <div> <input type="checkbox"/> GROSS CLOSURE CHARGE 150.00 </div> <div> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="2">TOTAL AMOUNTS</th> </tr> <tr> <th>GROSS</th> <th>NET</th> </tr> </thead> <tbody> <tr> <td>1,583.25</td> <td>1,345.76</td> </tr> <tr> <td><i>26.91</i></td> <td><i>26.91</i></td> </tr> <tr> <td><i>Net</i></td> <td><i>1,318.85</i></td> </tr> </tbody> </table> </div> </div>						DATE	TOTAL X DFT	DATE	TOTAL X DFT	DATE	TOTAL X DFT	7/10/88	4 X 15.75					7/11/88	6.00-00					TOTAL AMOUNTS		GROSS	NET	1,583.25	1,345.76	<i>26.91</i>	<i>26.91</i>	<i>Net</i>	<i>1,318.85</i>																																			
DATE	TOTAL X DFT	DATE	TOTAL X DFT	DATE	TOTAL X DFT																																																															
7/10/88	4 X 15.75																																																																			
7/11/88	6.00-00																																																																			
TOTAL AMOUNTS																																																																				
GROSS	NET																																																																			
1,583.25	1,345.76																																																																			
<i>26.91</i>	<i>26.91</i>																																																																			
<i>Net</i>	<i>1,318.85</i>																																																																			
<div style="display: flex; justify-content: space-between;"> <div> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>NEWSPAPER</th> <th>ISSUE</th> <th>DATE</th> <th>PRICE</th> <th>EXT.</th> <th>TOTAL</th> </tr> </thead> <tbody> <tr> <td>NASR</td> <td>6-10-88</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>ISSUE</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>DATE</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>PRICE</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>EXT.</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>TOTAL</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table> </div> <div> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>BRAND</th> <th>REV. DATE</th> <th>REV. NO.</th> <th>REV. DATE</th> <th>REV. NO.</th> </tr> </thead> <tbody> <tr> <td>HAR</td> <td>7-18-88</td> <td>7-18-88</td> <td></td> <td></td> </tr> <tr> <td></td> <td>7-18-88</td> <td></td> <td></td> <td></td> </tr> </tbody> </table> </div> <div> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>CREDIT</th> <th>DATE PAID</th> <th>CHECK #</th> </tr> </thead> <tbody> <tr> <td></td> <td>7-14-88</td> <td>00387</td> </tr> </tbody> </table> </div> <div> <div style="text-align: center;"> </div> </div> </div>						NEWSPAPER	ISSUE	DATE	PRICE	EXT.	TOTAL	NASR	6-10-88					ISSUE						DATE						PRICE						EXT.						TOTAL						BRAND	REV. DATE	REV. NO.	REV. DATE	REV. NO.	HAR	7-18-88	7-18-88				7-18-88				CREDIT	DATE PAID	CHECK #		7-14-88	00387
NEWSPAPER	ISSUE	DATE	PRICE	EXT.	TOTAL																																																															
NASR	6-10-88																																																																			
ISSUE																																																																				
DATE																																																																				
PRICE																																																																				
EXT.																																																																				
TOTAL																																																																				
BRAND	REV. DATE	REV. NO.	REV. DATE	REV. NO.																																																																
HAR	7-18-88	7-18-88																																																																		
	7-18-88																																																																			
CREDIT	DATE PAID	CHECK #																																																																		
	7-14-88	00387																																																																		
<div style="display: flex; justify-content: space-between;"> <div> <input type="checkbox"/> TOTAL COLLARS INDEXED </div> <div> <input type="checkbox"/> MAKE ALL CHECKS PAYABLE TO </div> <div> </div> </div>																																																																				
<div style="display: flex; justify-content: space-between;"> <div> FOR THIRD AVENUE, NEW YORK, N.Y. 10017 (212) 480-1200 </div> <div> 1,583.25 </div> <div> 1,345.76 </div> </div>																																																																				
NET DUE 1,345.76																																																																				


[illegible]

- with additional noise, grain or skew from the scanning process;

 Microbiological Associates A DIVISION OF Whittaker		4733 BETHESDA AVENUE BETHESDA, MARYLAND 20814 TELEPHONE (301) 654-3400 TWX 710-240		
		078-0018-1		
TO The Central Pur Bedrock Research, USA, INC. 110 West 59th Street New York, New York 10002 Attention: Dr. John Knudsen		INVOICE DATE: 11.13.75 DISCOUNT OF 5% ON NET AMOUNT IF PAID WITHIN 30 DAYS AFTER SHIPPING DATE, NET 30 DAYS		
SHIPPED TO		ORDER NO. 114625		
CUSTOMER NUMBER CUSTOMER P.O. NUMBER ORDER P.O. DATE DATE SHIPPED PACKING SLIP NO. ORDER NO.				
QUANTITY UNIT PRICE PRODUCT NUMBER DESCRIPTION UNIT PRICE AMOUNT				
This voucher represents refundable costs for period October 4, 1975 thru November 2, 1975				\$14,592.00
Fixed Fee				3,208.00
Total				\$14,800.00
114625 11/19				
NET AMOUNT \$14,800.00		TAX 0.00		TOTAL AMOUNT DUE \$14,800.00
FREIGHT CHARGES 0.00		RATE 0.00		AMOUNT 0.00
DUPLICATE INVOICE				(PRINT IN U.S. CURRENCY ONLY)

CTR CONTRACTS 011579

11231324

 Microbiological Associates A DIVISION OF Whittaker		4733 BETHESDA AVENUE BETHESDA, MARYLAND 20014 TELEPHONE: (301) 654-3400	
SOLD TO			
The Council for Tobacco Research, USA, Inc. 110 East 59th Street New York, New York 10022 Attn: Dr. John Kreishner SHIPPED TO		CTR-0028-7 ↑ <div style="border: 1px solid black; padding: 5px; text-align: center;"> PLEASE REFER TO OUR INVOICE NUMBER ON ALL PAYMENTS AND CORRESPONDENCE </div>	
		INVOICE DATE 11 13 75	
DISCOUNT OF % ON NET AMOUNT IF PAID WITHIN 20 DAYS AFTER SHIPPING DATE, NET 30 DAYS			
CUSTOMER NUMBER	CUSTOMER P.O. NUMBER	CURT. P.O. DATE NO. INVT. IN	DATE SHIPPED NO. INVT. IN
			PACKING SLIP NO.
QUANTITY	UNIT	PRODUCT NUMBER	DESCRIPTION
			This voucher represents reimbursable costs for period October 4, 1975 thru November 2, 1975 Fixed Fee Total
			14629
			\$6,994.16 310.00 \$7,304.16
NET AMOUNT	FREIGHT CHARGES	RATE	TAX
			AMOUNT
TOTAL AMOUNT DUE			
			\$7,304.16
DUPLICATE INVOICE			DEBIT IN U.S. CURRENCY ONLY

- with additional graphics and handwriting;

IN-001

Saffron Design
77 Namrata Bldg
Delhi, Delhi 400077

Bill To
Panchika Ranchhawa
27, DIF City, Central
Delhi, Delhi 40003

Ship To
Kavindra Mannan
284, Abdul Rehman
Mumbai, Bihar 40009

Invoice Date
29/01/2019

P.O.#
2439/2018

Due Date
29/04/2019

QTY	DESCRIPTION	UNIT PRICE	AMOUNT
1	Frontend design restructure	9,999.00	9,999.00
2	Custom icon package	875.00	1,850.00
3	Gandhi mouse pad	99.00	297.00
Subtotal			12,246.00
		VAT 6.0%	734.76

Invoice Total:

₹ 12,980.76

Terms & Conditions

Payment is due within 15 days.

State Bank of India
Account Number: 12345678
Routing Number: 08870563210

[illegible]

7 Comparison and Results

Tests were conducted from the perspective of a real end-user online on the official websites of the software products. Some of them were publicly free to use, others offered a trial version for a restricted time. The experiment was divided into 2 categories:

- **Field label**- for correctly identifying the field (e.g. recognize the field "Total Amount", "Total", "Total Due" as the total amount value on an invoice document, etc.)
 - In this case, the correctly classified fields were assessed as TRUE (>80% of the label characters were extracted accurately), the incorrectly classified ones, quasi the falsely extracted ones (e.g. "Recently", not the correct "Recipient"; "Sender", not the correct "Recipient") as well as the incorrectly classified as "not a field", quasi the missed fields were assessed as FALSE (Fig.9).
- **Field value** – for correctly extracting the information from the field (e.g. extract the value "2400.00 " from the "Total Amount" field)
 - Where in this case the correctly extracted field values were assessed as TRUE (>80% of the value characters were extracted accurately), the incorrectly extracted ones (e.g. from the field "Recipient" extracted value is "Jonn Oae", not the correct one "Jon Doe"), as well as the missed, non-extracted field values were assessed as FALSE (Fig.9).



Figure 9: Example of Field label/ Field value extraction [34].

Field label	Field value	Result
TRUE (e.g. "Recipient")	TRUE (e.g. "John Doe")	<u>TRUE POSITIVE (TP)</u>
TRUE (e.g. "Recipient")	FALSE(e.g. "Jonh Oae")	<u>FALSE POSITIVE (FP)</u>
FALSE (e.g. "Recently")	TRUE (e.g. "John Doe")	<u>FALSE NEGATIVE (FN)</u>
FALSE (e.g. "Recently")	FALSE (e.g. "Jonh Oae")	<u>FALSE NEGATIVE(FN)</u>

Table 3: Resulting values from the combination of field label and field value

The Values of the two sub experiments were then merged into one single value as shown in table 3, which can possess values: TRUE POSITIVE (TP), FALSE POSITIVE (FP) and FALSE NEGATIVE (FN). For FP were considered those of the values, which were with correctly extracted field label, but falsely extracted value. For FN were evaluated values, which were:

- with false field and correct value;
- with false field and false value extraction;

The three cumulative values TP, FP, FN were afterwards used to calculate the Precision ($P = TP / (TP + FP)$) and Recall ($R = TP / (TP + FN)$).

- In both experiments, for TP were considered only values, that match at least 80% of the "gold standard" ones and even the slightest errors in the content of the field label/field value below that were considered to be FN or FP. That was very important especially in the case of field value extraction because in many cases the extracted information was meaningless or false.
- Finally, the F1 Score ($F1 = 2 \times (P \times R / (P + R))$) was computed. F1 Score is typically used to combine Precision and Recall into a single harmonic mean value which is significantly more accurate than the typical arithmetic mean [25].

However, combining the Precision and Recall into a single value has its drawbacks, hence why it can conceal lower precision or recall values. That is why in our evaluation, before combining both values in the harmonic mean, we have firstly viewed them separately.

For each of the evaluated Tools (see Example Table 4 below), calculation of F1 Score was performed for every information field (*Sender Name; Sender Address; Receiver Name; Receiver Address; Invoice Number; Invoice Date; Total Amount*), so we have compared the results later and concluded which of them has the highest mean F1-Score, respectively the highest information extraction rate for the particular field.

Invoice Field	F1 Score
Sender Name	0.57
Sender Address	0.42
Receiver Name	0.64
Receiver Address	0.43
Invoice Number	0.81
Invoice Date	0.91
Total Amount	0.82

Table 4: Example Table for each of the fields

To get a better understanding of the performance, the errors that have occurred, were also analyzed and divided into 3 classes. Examples of each of them were given in table 7:

- **Wrong**– when the actual and extracted value are completely divergent; the Field label is correct but the field value is wrongly extracted or meaningless
- **OCR**- when the extracted characters differ from the actual ones (<80% similarity), only parts of the complete string are correctly extracted
- **Missing**- when no value is extracted, despite the presence of such, completely missed

We have also calculated the overall sum of the errors per tool out of all possible 350 errors.

7.1 Discussion

The Community version of Google Document AI, which was used for this experiment, works only with images in PDF format. In contrast to the other tools, which support various file formats, here jpg files and other document types are not supported. Invoice category had to be explicitly chosen as document type, otherwise, if the user chooses "General document" or "Form 1040", invoice features were absent. The OCR is very accurate and the majority of related errors were of type wrong.

At its core, Rossum does not support "Sender Address" and "Receiver Address" as values, as mentioned in the tables. However, the extraction results for Sender Name and Receiver Name were much higher and stable. The text had to be explicitly annotated- if not, Rossum wasn't able to read it. That was the case with "Invoice Number" and "Total Amount" fields. Furthermore, on some of the invoices containing the dollar sign- "\$", the outputted value was 5, which we have considered as an OCR type of error.

In Hypatos, users can choose between model types of documents before importing them. The invoice categories are EU, International Invoice, Receipt and Invoice community. The tool does not manage well pictures with noise or grain and then the OCR engine wasn't very accurate.

Parashift offers very precise extraction from colored and grainless invoices, as well as recognition of handwriting on the document. However, some of the Documents with lower resolution or noise were unreadable and unprocessable.

As well in Syphr, the black and white or lower resolution pictures weren't sufficiently recognized, as such only "Invoice number" and "Invoice Date" were extracted and processed accurately.

The case with Docparser was very specific. As a template-based tool, it counts on predefined rules for the extraction process. Before importing the 50 documents into the system, 10 extraction rules for every field were created. Unfortunately, even with annotation rules performance was low especially the low extraction rate at "Sender Address" and "Receiver Address" fields. The tool "suffers" from mediocre OCR performance as well.

As we can recognize from table 5, all of the tested tools have shown higher precision than recall (recall is underlined) with the highest levels of precision achieved by Google, Docparser and Parashift. Unfortunately, in the case of Google and Doc-parser, higher Precision values don't transfer in higher overall score, due to the lower Recall. Parashift shows the most consistent performance among all the invoice fields and gets the highest recall in 5 out of the 7 fields and very competitive precision. Syphr gets the other two- for "Invoice Number", "Invoice Date" and shows very good achievement at

"Total Amount", but lacks enough sufficient recall in the other 4 fields- "Sender Name", "Sender Address", "Recipient Name", "Recipient Address".

The "perfect" precision benchmarks of some of the tools can lead to false security because almost no false positive values are generated. If one correct answer can be found, then only that answer can be returned and the number of false positives would be nonexistent. But the volume of false negatives was at a much higher rate, thus heavily affecting the recall benchmark [24].

Both values were afterwards combined in the F1 score value, shown in table 6. At first glance, it becomes obvious that the "Sender Name", "Sender Address" along with "Recipient Name", "Recipient Address" were the most ambiguous for extraction fields, because of their complexity and the diversity of information included inside of them- company/personal name, the full address with postcode, city, state and street of the participant.

General flaw, that almost all of the tools had regarding "Sender" and "Receiver" were switching content. There were cases in which, the system recognized the address of the Sender as the address of the Receiver or reverse. Similar was the case with the names of both parties. Such cases were considered as errors of type "Wrong".

Another common problem was with "P.O" (purchase order number) and "Invoice Number". Systems had frequently mistaken both of the fields and delivered false extraction.

Invoice Field	Google		Rossum		Hypatos		Parashift		Sypht		Docparser	
	P	R	P	R	P	R	P	R	P	R	P	R
Sender Name	<i>1</i>	0,6	0,94	0,7	0,87	0,75	0,95	<u>0,91</u>	0,81	0,72	0,77	0,3
Sender Address	0,96	0,57	0	0	0,84	0,75	<i>1</i>	<u>0,88</u>	0,89	0,76	<i>1</i>	0,14
Receiver Name	0,95	0,47	0,97	0,81	0,78	0,6	<i>1</i>	<u>0,88</u>	0,89	0,55	0,92	0,24
Receiver Address	<i>1</i>	0,46	0	0	0,73	0,6	0,97	<u>0,85</u>	0,81	0,39	0,9	0,2
Invoice Number	<i>1</i>	0,88	0,89	0,76	0,92	0,74	<i>1</i>	0,88	0,97	<u>0,93</u>	<i>1</i>	0,22
Invoice Date	<i>1</i>	0,9	0,95	0,95	0,96	0,65	0,97	0,93	0,97	<u>0,97</u>	<i>1</i>	0,4
Total Amount	0,97	0,93	0,89	0,97	0,88	0,86	<i>1</i>	<u>0,98</u>	0,97	0,89	<i>1</i>	0,68

Table 5: Precision and Recall

(Legend: Precision- highest score per field in bold and italic; Recall- highest score per field in bold and underlined)

Invoice Field	Google AI F1 Score	Rossum F1 Score	Hypatos F1 Score	Parashift F1 Score	Sypht F1 Score	Docparser F1 Score
Sender Name	0,75	0,8	0,8	0,93	0,76	0,43
Sender Address	0,71	0	0,79	0,93	0,82	0,24
Receiver Name	0,64	0,88	0,68	0,93	0,68	0,38
Receiver Address	0,63	0	0,66	0,91	0,52	0,33
Invoice Number	0,93	0,82	0,82	0,93	0,94	0,36
Invoice Date	0,94	0,95	0,78	0,95	0,97	0,57
Total Amount	0,93	0,93	0,87	0,98	0,93	0,8

Table 6: F1 Score

(Legend: F1 Score- highest score per field in bold)

7.2 Error Analysis

After performing the benchmark, the final results of the error classification are shown in Table 8. It is firstly evident that the majority of the tools except Google AI, made mainly mistakes of type "missing", quasi completely ignoring the content and label of the field.

The second most common errors are of type "wrong"- when for example correct "OCRred" value was returned but in most cases, this value was from another field. The other type of "wrong" is when Field extraction is correct but the field label is wrongly extracted or meaningless.

The "OCR" errors were the rarest, ranging from 5% in Google AI to almost 25 % in Hypatos.

Error Type	Correct Value (Gold Standard)	Value returned
Wrong	Expected: Turnpike Designs Co.	Toronto, Canada
OCR	Expected: Turnpike Designs Co.	Tupmiks Dacia Ca.
Missing	Expected: Turnpike Designs Co.	-

Table 7: Examples of each error type

Error Type	Google AI	Rossum	Hypatos	Parashift	Sypht	Docparser
Wrong	94%	6%	33%	6%	27%	21%
OCR	5%	8%	25%	8%	18%	4%
Missing	1%	86%	42%	86%	55%	75%
Total Sum Errors	111	151	128	37	103	246
Total Error Rate	31%	43%	36%	10%	29%	70%

Table 8: Error Type benchmark

(Legend: highest error rate per type in bold)

8 Related Work

So far, significant progress has been accomplished in the field of data extraction. The companies struggled with document processing automatization for decades and the attempts for optimizing the process date back since the nineties.

Such generic system for processing invoices [35], presented by T. A. Bayer et al. , extracts the requested fields automatically from documents with different domains and layouts. The architecture consists of two main components- an OCR tool for domain knowledge, which is used for converting the image into layout structure and a domain knowledge language FRESCO (Frame Representation Language for Structured Documents), used for describing the desired items.

When it comes to commercially available tools, two papers come to the forefront – the first [36] is by Holt et al, which presents SyphT- scalable context-driven solution to document field extraction. It combines heuristic filtering, OCR and supervised machine learning and manages to work with skewed and lower resolution images. SyphT took part in our performance benchmark as well. The second [37] by Schuster et al presents Intellix, which is deployed in the document management tool DocuWare, was trained with user-generated training examples and excludes rule-based information extraction, which demands a lot of training data. The system relies on an external OCR system.

In relation to the subject of deep learning and neuronal networks, we have analyzed multiple papers. The first work [38] by Raoui-Outach et al. aims at understanding images of sale receipts, in particular fields like store brand, purchased products, prices, etc. with the help of Deep Convolutional Neural Networks (DCNNs) and classical image and text processing. The paper [39] by Tata et al. proposes a novel approach for extracting information based on a neuronal network from images of form-like documents (invoices, purchase orders) and generates extraction candidates, that learn a dense representation based on the neighbouring words. The system renders each document into an image and uses cloud OCR service to extract the text.

Regarding tables in invoices, [40] by Paliwal et al. proposes a deep learning model for table detection and tabular data extraction from images. Prior approaches tried to solve these tasks independently from each other. The proposed model exploits the interdependence between these tasks to divide the table and columns. Next, rule-based row extraction from the identified tabular sub-regions is performed.

We identify another emerging topic in the scientific literature in the last couple of years, namely Few-shot learning/One-shot learning [41]. Its

minimalistic nature aims to overcome the requirement of large-scale datasets for training the machine learning models and it can gain information from just a few training examples, sometimes even less than ten. A paper [42] by Esser et al. presents a solution for the Small office /home office(SOHO) users, which typically work with a smaller quantity of documents and data. Their approach is based on template document detection, which manages to identify similar documents with the help of textual and layout-based features together with the text search engine library Lucene [43]. The similar documents are subsequently used as input in the extraction algorithm to generate accurate results for each component.

Sunder et al. [44] attempt to deal with the one-shot information extraction concept from images of invoices, bills, etc. Their approach is in two stages. In the first step, a pre-trained neuronal network is used to read the information in the training images and store it in a database. In the second, deductive learning is used to learn the extraction algorithm- more specifically a resource-bounded meta-interpreter creates proofs with the help of the training examples via logical deduction.

9 Conclusion

This work presents an independent benchmark study of existing data extraction tools, publicly available for use on the Internet. The task consisted of extraction of 5 key-value fields, contained on 50 invoice images with different layouts and fonts.

We have concluded that:

- From all of the tested tools, Parashift scored the highest score in our benchmark regarding the measure precision, recall and F1 score and made the lowest amount of errors
- For the majority, the most common type of errors was missing the field and its value
- Grained/Skewed and lower resolution document images tend to be problematic for the processing software
- Future works and tools in the area have to manage working with a smaller set of training images to attract more potential customers

In general, the machine learning-based tools scored higher than the template-based. Template-based solutions make sense only if the customer has a limited quantity of vendors and the type of processed documents is confined. The use of such tools is limited when the layout is diverse and no prior annotation of such type of document template is available. Because of that, they can find use only in situations, where document layouts are limited and the user has previously annotated the training examples.

Hence, machine learning-based solutions offer much more versatility when the document input is higher and the layout is unknown or variable.

References

- [1] Concur, "Supplier Invoice Benchmark Report: Exploring the AP Landscape in the UK," 2015.
- [2] S. Marinai and A. Dengel, *Document analysis systems VI: 6th international workshop, DAS 2004, Florence, Italy, September 8-10, 2004 proceedings*. Berlin, New York: Springer, 2004.
- [3] C. Dilmegani, *Invoice, the Critical Business Document, Explained [2021]*. [Online]. Available: <https://research.aimultiple.com/invoice/> (accessed: Feb. 26 2021).
- [4] *Definition information extraction (IE)*. [Online]. Available: <https://whatis.techtarget.com/definition/information-extraction-IE> (accessed: 26.09.20).
- [5] B. Lutkevich, *Definition natural language processing (NLP)*. [Online]. Available: https://searchenterpriseai.techtarget.com/definition/natural-language-processing-NLP?_gl=1*126p0vv*_ga*MTA3MTQyNDUwNy4xNjE0ODEwNzIx*_ga_RRBYR9CGB9*MTYxNDgxMDcyMC4xLjAuMTYxNDgxMDcyMC4w&_ga=2.190466027.1943557216.1614810721-1071424507.1614810721 (accessed: 26.09.20).
- [6] Optiform, *Structured and Unstructured Documents - What is the Difference?*, 2016. Accessed: 23.10.20. [Online]. Available: <https://www.optiform.com/news/structured-unstructured-documents/>
- [7] A. K. Y. Ommi, *1.1 Introduction to Data and Information*. [Online]. Available: <https://i0.wp.com/www.mycloudwiki.com/wp-content/uploads/2016/06/1-1-1024x392.jpg> (accessed: 11.11.20).
- [8] B. Koch, "The e-invoicing journey 2019-2025," 2019.
- [9] IBM Corporation, *The Future of EDI: An IBM point of view*, 2020. Accessed: Feb. 28 2021. [Online]. Available: <https://www.ibm.com/downloads/cas/WQB6NBJ7>
- [10] Mittelstand 4.0- Agentur, *Elektronische Rechnungsabwicklung und Archivierung: Fakten aus der deutschen Unternehmenspraxis 2017*, 2017.
- [11] C. Dilmegani, *Invoice Capture: Guide to most firm's first AI purchase in 2021*. [Online]. Available: <https://research.aimultiple.com/invoice-capture/#what-is-invoice-capture-> (accessed: 17.01.21).
- [12] M. Rüfenacht, *Template-Based OCR Versus Machine Learning-Based OCR*, 2020. Accessed: 12.12.20. [Online]. Available: <https://parashift.io/en/template-based-ocr-versus-machine-learning-based-ocr/>
- [13] A. Jnagal, *Templates Vs Machine Learning OCR*, 2018. Accessed: Feb. 28 2021. [Online]. Available: https://www.infrd.ai/blog/templates-vs-machine-learning-ocr?utm_campaign=UTM_Tracking_Insights_Page&utm_source=Backlinks&utm_content=Blogs%20-%20Templates%20vs%20Machine%20Learning%20OCR
- [14] B. Settles, *Active Learning Literature Survey*: University of Wisconsin-Madison Department of Computer Sciences, 2009.

- [15] P. Klügl, *Context-specific Consistencies in Information Extraction: Rule-based and Probabilistic Approaches*. Zugl.: Würzburg, Univ., Diss., 2014. Würzburg: Würzburg Univ. Press, 2015. [Online]. Available: <http://nbn-resolving.de/urn:nbn:de:bvb:20-opus-108352>
- [16] S.Sarawagi, "Information Extraction," *FNT in Databases*, vol. 1, no. 3, pp. 261–377, 2007, doi: 10.1561/19000000003.
- [17] J.Rausch, O.Martinez, F.Bissig, C.Zhang, S.Feuerriegel, *DocParser: Hierarchical Structure Parsing of Document Renderings*, 2019.
- [18] R. B. Palm, O. Winther, and F. Laws, "CloudScan - A configuration-free invoice analysis system using recurrent neural networks," Aug. 2017. [Online]. Available: <https://arxiv.org/pdf/1708.07403>
- [19] T. B. K. Baierer, *hOCR - OCR Workflow and Output embedded in HTML*. Living Standard, 2020. [Online]. Available: <http://kba.cloud/hocr-spec/1.2/> (accessed: Feb. 24 2021).
- [20] *What is n-gram*. [Online]. Available: <https://www.definitions.net/definition/n-gram> (accessed: 12.10.20).
- [21] H. W. Kuhn, *The Hungarian method for the assignment problem*, 1955.
- [22] G. K.Holman, *Universal business language v2.0*. [Online]. Available: <https://docs.oasis-open.org/ubl/os-UBL-2.0/UBL-2.0.html> (accessed: 11.11.20).
- [23] K.Bengtsson, *Universal Business Language*. [Online]. Available: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=ubl (accessed: 21.10.20).
- [24] L.Derczynski, *Complementarity, F-score, and NLP Evaluation*, 2016.
- [25] C. D. Manning, P. Raghavan, and H. Schütze, "Evaluation in information retrieval," in *Introduction to information retrieval*, C. D. Manning, P. Raghavan, and H. Schütze, Eds., New York: Cambridge University Press, 2008, pp. 139–161.
- [26] K.P. Shung, *Accuracy, Precision, Recall or F1?* [Online]. Available: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9> (accessed: 22.12.20).
- [27] S.Tata, *Extracting Structured Data from Templatic Documents*, 2020. Accessed: 05.10.20. [Online]. Available: <https://ai.googleblog.com/2020/06/extracting-structured-data-from.html>
- [28] FreshBooks, *What Are Different Types of Invoices? | Small Business Invoicing Guide*. [Online]. Available: <https://www.freshbooks.com/hub/invoicing/types-of-invoices> (accessed: 24.12.20).
- [29] D.Mudliar, "Different Types of Invoices in Accounting," 2020. [Online]. Available: <https://www.billbooks.com/blog/types-of-invoices/>
- [30] M. Dankowska, *What Information Does An EU Invoice Need?* [Online]. Available: <https://www.vertabelo.com/blog/what-information-does-an-eu-invoice-need-advice-from-a-tax-expert/> (accessed: 11.01.21).
- [31] Taxation and Customs Union, *VAT invoicing rules*. [Online]. Available: https://ec.europa.eu/taxation_customs/business/vat/eu-vat-rules-topic/vat-invoicing-rules_en (accessed: 15.10.20).

- [32] Vertabelo. Accessed: 20.11.20. [Online]. Available: <https://www.vertabelo.com/blog/what-information-does-an-eu-invoice-need-advice-from-a-tax-expert/vertabelo-invoice.jpg>
- [33] A.W.Harley,A.Ufkes, K.Derpanis, *The RVL-CDIP (Ryerson Vision Lab Complex Document Information Processing) dataset*. [Online]. Available: <http://www.cs.cmu.edu/~aharley/rvl-cdip/> (accessed: 22.11.20).
- [34] *Invoice-dataset*. [Online]. Available: <https://github.com/SouravG94/invoice-dataset> (accessed: 22.11.20).
- [35] T. A. Bayer, H. U. Mogg-Schneider, *A Generic System for Processing Invoices*, 1997.
- [36] A.C. X.Holt, *Extracting structured data from invoices*: Proceedings of the Australasian Language Technology Association Workshop 2018, 2018.
- [37] D.Schuster, K.Muthmann, D.Esser, A.Schill, M.Berger, C.Weidling, K.Aliyev, A.Hofmeier, *Intellix - End-User Trained Information Extraction for Document Archiving*, 2013. [Online]. Available: <https://ieeexplore.ieee.org/document/6628593>
- [38] R. Raoui-Outach, C. Million-Rousseau, A. Benoit, P.Lambert, *Deep Learning for automatic sale receipt understanding*. Piscataway, NJ: IEEE, 2017.
- [39] B.Majumder,N.Potti, S.Tata ,J.B. Wendt,Q.Zhao,M.Najork, *Representation Learning for Information Extraction from Form-like Documents*, 2020.
- [40] S.Paliwal, Vishwanath D, R.Rahul, M.Sharma, L.Vig, "TableNet: Deep Learning Model for End-to-end Table Detection and Tabular Data Extraction from Scanned Document Images," TCS Research, New Delhi, India, 2019.
- [41] I.A.Ozsubasi, *Few-Shot Learning (FSL): What it is & its Applications*. [Online]. Available: <https://research.aimultiple.com/few-shot-learning/> (accessed: 26.11.20).
- [42] D.Esser,D.Schuster,K.Muthmann,A.Schill, *Few-exemplar Information Extraction for Business Documents*, 2014.
- [43] Apache, *Apache Lucene Core*. [Online]. Available: <https://lucene.apache.org/core/> (accessed: 12.11.20).
- [44] V.Sunder, A.Srinivasan, Vig, Shroff, R.Rahul, *One-shot Information Extraction from Document Images using Neuro-Deductive Program Synthesis*, 2019.