







# Information Extraction from Business Documents

## A Case Study

Martin Geletka , Mikuláš Bankovič , Dávid Meluš , Šárka Ščavnická ,  
Michal Štefánik , and Petr Sojka 

Faculty of Informatics, Masaryk University  
Botanická 68a, 602 00 Brno, Czech Republic  
456576@mail.muni.cz

**Abstract.** Document AI is a relatively new research topic that refers to techniques for automatically reading, understanding, and analyzing business documents. Nowadays, many companies extract data from business documents through manual efforts that are time-consuming and expensive, requiring manual customization or configuration. This paper describes techniques to address these problems, apply them to real-world data, and implement them to an end-to-end solution for automatic information extraction from business documents.

**Keywords:** OCR, Multi-modal learning, Information extraction, Transformers, Structured Documents

## 1 Introduction

Information extraction typically consists of two consecutive steps. Firstly text detection and recognition are run to obtain text representation of the input document. Secondly, we extract the information from the received text. In our paper, we present a multi-modal approach to information extraction, which extracts information not only from text alone but integrates all three modalities: text, position, and image, to obtain the best results.

## 2 OCR frameworks

An essential step at the start of the business document pipeline is finding text blocks and their positions on the page. For scanned documents, OCR (Optical Character Recognition) frameworks are needed.

There are recent OCR frameworks based on deep learning: we describe the models and steps used in these frameworks. We focus on two main steps in OCR pipelines: text detection and text recognition. We discuss the importance of customization and fine-tuning the models included. Specifically, we compare frameworks: Doctr [11], EasyOCR [6], and Tesseract [13] and their ability to be customized and fine-tuned for document understanding in the Czech language.

Text detection models aim to output areas containing text. Most of these models are trained in languages based on Latin script. Therefore, we assume the performance does not suffer without training or fine-tuning the models for the Czech language and is sufficient to perform document understanding. Scene text detection is an active area of research and can be easily re-used in document understanding.

Text recognition models aim to extract text from the bounding boxes generated by text detection models. These models need to be fine-tuned for the specific language vocabulary. We use Differentiable Binarization Net (DBNet) [9] as it is available in both EasyOCR and Doctr<sup>1</sup>. We can compare text recognition models in an end-to-end pipeline by unifying text detection architecture. EasyOCR has additionally available CRAFT [3] model.

## 2.1 Models

This subsection briefly introduces different model architectures we are training or re-using.

**DBNet** Text detection model, that proposes Differentiable Binarization. The model produces a segmentation map alongside the proposed threshold for binarization. The threshold map solves post-processing and improves metric performance and speed. Model implementations can differ in Convolutional Neural Network (CNN) backbone<sup>2</sup>. The most common is original *vgg-16*, *resnet18*, deeper *resnet50*.

**Convolutional Recurrent Neural Network (CRNN)**[12] Text recognition model that combines the strengths of image feature extraction of CNN followed by the sequential processing of RNN. CNN's local patch processing ensures that columns from the feature space correspond to column patches in an original image. Locality preserving allows sequential processing in RNN. The most common RNN layers are GRU or LSTM to model long-term dependencies.

**MASTER: Multi-Aspect Non-local Network for Scene Text Recognition** [10] Text recognition model that introduces Multi-Aspect Global Context Attention (GCAttention) based encoder module and a transformer-based decoder module.

**Vision Transformer for Fast and Efficient Scene Text Recognition (ViTSTR)** [2] Text recognition model that follows transformers architecture with self-attention mechanism and multi-headed attention. This model emphasizes speed and efficiency in a single-stage encoder step based on vision transformer (ViT) [4].

<sup>1</sup> Implementation of DBNet in EasyOCR and Doctr differ in backbone and weights

<sup>2</sup> Implementation of CRNN in EasyOCR, Doctr, and Tesseract differ in backbone and weights.

## 2.2 Tesseract

Tesseract is a well-known OCR framework that is considered to be the open-source baseline. Since version 4.0.0, Tesseract has used the CRNN architecture with the LSTM network for text recognition. Text detection is still performed using multiple steps: component analysis, contour detection, and lines detection. Traditional text detection causes Tesseract to be more prone to preprocessing techniques. In order to get better OCR results, improvement of the quality of the image is needed.

## 2.3 EasyOCR

EasyOCR framework offers only the CRNN model as a baseline. The CRNN is pre-trained on an English text and it combines Convolutional Neural Networks and Recurrent Neural Networks. EasyOCR's pre-, mid- and post-processing are parametrized and customizable. The parameters for text bounding box merging can change granularity from lines to words, and slope parameters adjust how much rotation is allowed in text bounding boxes. However, training scripts need to be better documented.

## 2.4 Doctr

Doctr framework is very recent and contains CRNN, MASTER and ViTSTR model architectures. We use this framework for custom training as it contains well documented curated repository with novel architectures.

# 3 Multi-modal Transformers overview

In this section, we will introduce the multi-modal models, which use additional modalities, such as position and image, to maximize the performance of information extraction tasks. In more detail, we will discuss the Layout Language Model (LayoutLM) Family developed by Microsoft Corporation. We will describe the three generations of the LayoutLM models together with its cross-lingual version LayoutXLM. We will also list related work by other research groups to obtain the whole picture of Multi-modal models.

## 3.1 LayoutLM family

The first model of the LayoutLM family has introduced already in December 2019. [17] Architecture of this first model was a quite simple extension of the Vanilla Transformer model. Instead of simple WordPiece tokens, this model takes on input also individual positions of the bounding boxes of corresponding tokens. The context-aware embeddings from the Transformer models are then concatenated with the document representation from a pre-trained Vision Neural Network.

The main difference between the first and second versions of the LayoutLM model is that the LayoutLMv2 takes the image representation on the input of Transformer models and therefore is able to train attention between all three modalities at once.

The improvement in the third version of the model was to use a domain-specific model for the Vision Embeddings. The model used a Document Image Transformer pre-trained as Auto-Encoder on IIT-CDIP, a dataset that includes 42 million document images [14].

The main training objective for these models is still a variation of the Mask Language Modelling, which aims to predict masked text tokens based on their position and surrounding text and image context information. The models also use additional pre-training tasks, which can be found in the published paper of all three models. [17,16,5]

All three versions of the models were trained on IIT-CDIP Test Collection [14]. The collection contains 11 million scanned documents. The dataset consists of documents from the state’s lawsuit against the tobacco industry and extracted text provided by the OCR system in the 1990s.

### 3.2 Multi-lingual models

The LayoutLM family is extended to LayoutXLM (Layout Cross-Lingual Language Model) to address the problem of information extraction from documents from multiple languages. [18] This model architecture and pre-training are inherited from LayoutLMv2 but pre-trained and evaluated on different datasets. The lack of some extended multi-lingual scanned document dataset forced the researchers to crawl the web for digital-born multi-lingual documents. Scrapped were then parsed with a PDF parser and filter from records containing less than 200 words or containing more than one language (identified by language detector from the BlingFire<sup>3</sup>). The dataset was then enriched by sampling from scanned English documents from IIT-CDIP Test Collection. The final dataset contained more than 30M documents in 53 languages (including Czech and Slovak).

The only other model we found pre-trained on the multi-lingual dataset is LiLT [15]. This model offers the option to divide the text representation and layout into two separate models, which can be pre-trained separately and only fine-tuned together. Therefore in fine-tuning, one can use any pre-trained language model such as XLM-RoBERTa and then only merge it with Layout Transformer and fine-tune them together. For more information about how the textual model is separated, we refer the reader to the original paper [15].

### 3.3 Other related work

Extensive research exists in multi-modal transformers, but to our knowledge, only LayoutLM and LiLt also offer multi-lingual models. In this section, we provide a short overview of conducted work in this area and for further

<sup>3</sup> <https://github.com/microsoft/BlingFire>

information, refer the reader to the original papers. These models also belong to related multi-modal models:

- **FormNet** – model from Google AI Research, which proposes two new mechanisms called Rich Attention and Super-Tokens. Rich Attention leverages the spatial relationships between tokens to calculate a more structurally meaningful attention score and Super-Tokens for each word in a form by embedding representations from their neighboring tokens through graph convolutions. [7]
- **DocFormer** – model from Amazon AI Research Group, which offers another type of multi-modal attention implemented through residual connections and contributes to additional pre-training tasks. [1]
- **SelfDoc** – from Brandeis University and Adobe Research group, which main difference to LayoutLM family models is that it adopts semantically meaningful components (e.g., text block, heading, figure) as the model input instead of WordPiece tokens. [8]

## 4 Experiments

This section will describe the dataset used for training our models, followed by a description of the individual experiments and a comparison of the trained models.

### 4.1 Dataset description

We perform a collection of documents, as no Czech documents dataset of a sufficient volume or quality is available. Our collection is performed by querying and automated downloading of the document-format file results from a publicly-available data-sharing platform [uloz.to](https://uloz.to). We query for keywords associated with common categories of office documents, such as “faktura”, “smlouva” or “doklad”. Such-collected, categorized documents are then manually cleaned, resulting in a collection of 6,849 invoice images that we annotate for chosen entity types.

We obtained languages of crawled documents by applying publicly available language detection tool<sup>4</sup> on the output of the Tesseract OCR engine.

In Figure 3, we can see that the final dataset contains documents mainly in Czech, Slovak, and Polish but also small amounts of English and Slovenian invoices.

The annotation process consists of (i) selecting a bounding box (BBox) that *separates* a position of the entity within the document visual and (ii) assigning a category to such BBox out of a predefined set of entity labels<sup>5</sup>. In Figure 1, we

<sup>4</sup> Tool is available at <https://github.com/Mimino666/langdetect>

<sup>5</sup> The entity types are chosen to allow an automated payment of the detected invoice based on the extracted information.

can see an example of an invoice with annotated entities with corresponding bounding boxes.

The annotation process yields a total of 39,670 entity annotations, ranging from 10 annotations for the rarest category (specific symbol) to 8,359 annotations for the most common category (total amount).

## 4.2 Born-digital dataset

By the same procedure, we receive a collection of 788 pdf documents consisting of born-digital invoices and scanned invoices. Firstly, we clean the dataset by removing scanned documents. As a criterion for scan identification, we use the average size of the page, dimensions of images, and characters. More precisely, we reject documents with an average page size greater than 500,000B or documents containing images with dimensions larger than the dimensions of the pdf. Further, we inspect characters, and documents containing unknown characters are removed. Through this process, we obtain 485 documents that are processed and used for fine-tuning Doctr and EasyOCR models.

We use pdfminer program to extract words (labels) and the corresponding images, resulting in a dataset of 687,241 samples. We train OCR models on this dataset with a train-validation-test split on unique words (60/20/20).

## 4.3 Results OCR


In Table 1, we compare pre-trained Tesseract, EasyOCR, and Doctr CRNN models with our trained models Doctr MASTER and Doctr ViTSTR. These models are not available pre-trained, and our training of the Doctr CRNN model was unsuccessful due to an error in the library. We compared the exact match, partial match, and elapsed time. The exact match is a 1– word error rate. A partial match is 1 if the ground truth starts with the full prediction; otherwise, 0.

Doctr MASTER performed the best from our tested models with 2% word error rate. However, it is a magnitude slower than Tesseract. Doctr ViTSTR is the faster-trained model; however, its performance is insufficient for commercial use. Pre-trained EasyOCR model is faster and has better performance than Doctr ViTSTR.

## 4.4 Results

In this section, we will compare the performance of text-only models with the same-sized LayoutLM models. We also compared models with two different pretraining datasets, i.e. pretrained only on English data and pretrained on multiple languages, including Czech and Slovak.

As a representative of the text-based model pretrained on English, we chose RoBERTa Large model and pretrained on the multilingual dataset BERT Vase Multilingual Cased, XLM RoBERTa Base, and XLM RoBERTa Large. From the

|  <b>Dodavatel:</b><br><b>VIDOX s.r.o.</b><br>U Poráků 511, Horní Brána<br>381 01 Český Krumlov<br>Česká republika<br>IČ: 25160168<br>DIČ: CZ25160168<br><b>Stavební divize:</b><br><b>Vodárenská 1091/II, 379 01 Třeboň</b><br><small>Dodávatel je registrován pod zvláštní značkou úřadu<br/>         C. vložka 6919 za dne 24.03.1997 u Krajského soudu<br/>         v Českých Budějovicích.</small><br>Úhrada: Na bankovní účet<br>Banka: Komerční banka a.s. Český<br>Číslo účtu: 4255580287/0100<br>IBAN: CZ5401000000004255580287<br>SWIFT: KOMB CZPPXXX   | Variabilní symbol (uvádějte při platbě): <b>300008616</b><br>Strana č. 1<br><b>Faktura - daňový doklad č.:</b> <b>300008616</b><br><b>Odběratel:</b> Zákaznické číslo: 107636<br><b>Husitské muzeum v Táboře</b><br>nám. Mikuláše z Husi 44/5<br>390 01 Tábor<br>IČ: 00072486<br>DIČ: CZ00072486 |                         |                         |                     |            |                   |                   |   |  |  |  |  |  |  |  |
|---|--|-------------------------|-------------------------|---------------------|------------|-------------------|-------------------|---|--|--|--|--|--|--|--|
|   | Datum: 10-08-2016<br>Č. j.: HM-CP/604/2016<br>listy: 1 přílohy: 1<br>zprávy: 1<br>příl. zprávy: 1  |                         |                         |                     |            |                   |                   |   |  |  |  |  |  |  |  |
| Středisko: 300 Akce:<br>Datum vystavení dokladu: 5.8.2016 Datum splatnosti: 4.9.2016<br>Datum zdanitelného plnění: 31.7.2016<br>Místo plnění: CZ  |  |                         |                         |                     |            |                   |                   |   |  |  |  |  |  |  |  |
| <table border="1"> <thead> <tr> <th>Předmět zdanitelného plnění</th> <th>Množství Jj.</th> <th>Cena za jedn. v CZK bez</th> <th>Cena celkem bez DPH</th> <th>Sazba DPH</th> <th>Částka DPH</th> <th>Cena celkem s DPH</th> </tr> </thead> <tbody> <tr> <td colspan="7">           Fakturuje mě Vám provedené stavební práce na stavební zakázce "Stavební úpravy a přístavba č. p. 2073 Tábor" dle uzavřené SOD č. objednávky S/2016/PO za období 1.7.2016 - 31. 7. 2016 a zjišťovacího protokolu č. 5, který tvoří nedílnou součást této faktury<br/>           Částka ve výši: 1 455 507,17 0% 0,00 1 455 507,17<br/>           Veřejná zakázka - TENDERMARKET pod ev. č. T004/15V/00048992<br/>           Upozornění: 1) Daň odvede zákazník         </td> </tr> </tbody> </table> | Předmět zdanitelného plnění  | Množství Jj.            | Cena za jedn. v CZK bez | Cena celkem bez DPH | Sazba DPH  | Částka DPH        | Cena celkem s DPH | Fakturuje mě Vám provedené stavební práce na stavební zakázce "Stavební úpravy a přístavba č. p. 2073 Tábor" dle uzavřené SOD č. objednávky S/2016/PO za období 1.7.2016 - 31. 7. 2016 a zjišťovacího protokolu č. 5, který tvoří nedílnou součást této faktury<br>Částka ve výši: 1 455 507,17 0% 0,00 1 455 507,17<br>Veřejná zakázka - TENDERMARKET pod ev. č. T004/15V/00048992<br>Upozornění: 1) Daň odvede zákazník |  |  |  |  |  |  |  |
| Předmět zdanitelného plnění   | Množství Jj.   | Cena za jedn. v CZK bez | Cena celkem bez DPH     | Sazba DPH           | Částka DPH | Cena celkem s DPH |                   |   |  |  |  |  |  |  |  |
| Fakturuje mě Vám provedené stavební práce na stavební zakázce "Stavební úpravy a přístavba č. p. 2073 Tábor" dle uzavřené SOD č. objednávky S/2016/PO za období 1.7.2016 - 31. 7. 2016 a zjišťovacího protokolu č. 5, který tvoří nedílnou součást této faktury<br>Částka ve výši: 1 455 507,17 0% 0,00 1 455 507,17<br>Veřejná zakázka - TENDERMARKET pod ev. č. T004/15V/00048992<br>Upozornění: 1) Daň odvede zákazník   |  |                         |                         |                     |            |                   |                   |   |  |  |  |  |  |  |  |

|                        | Částky v CZK |      |                     |
|------------------------|--------------|------|---------------------|
|                        | Bez DPH      | DPH  | Celkem              |
| 0 %                    | 1 455 507,17 | 0,00 | 1 455 507,17        |
| Celkem                 | 1 455 507,17 | 0,00 | 1 455 507,17        |
| Zaokrouhlení           |              |      | 0,00                |
| Na zálohách zapláceno  |              |      | 0,00                |
| <b>Částka k úhradě</b> |              |      | <b>1 455 507,17</b> |

Základem pro výpočet daně je částka "Bez DPH".

Vystavil(a): Kateřina Hloušková

Převzal(a), dne: 10.8.2016



Vytvářeno v systému ABRAC3

Telefon: +420384721357

Fax: +420384721357

E-mail: katerina.hlouskova@vidox.cz

Mobilní telefon:

WWW: www.vidox.cz

Fig. 1: Example of annotated invoice

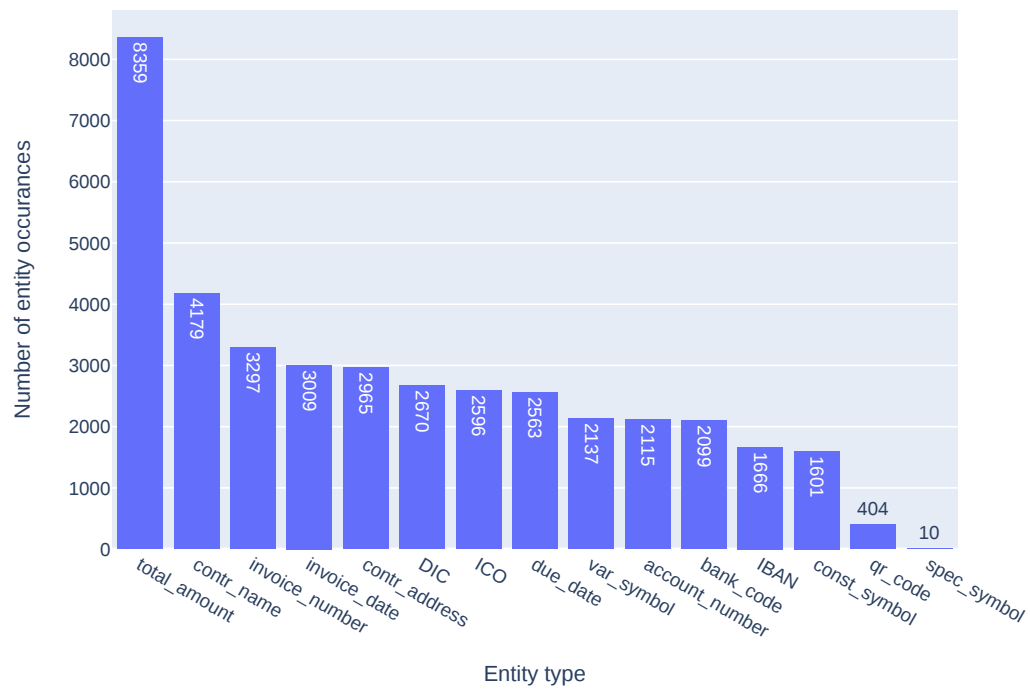


Fig. 2: Number of individual entity types across the whole dataset.

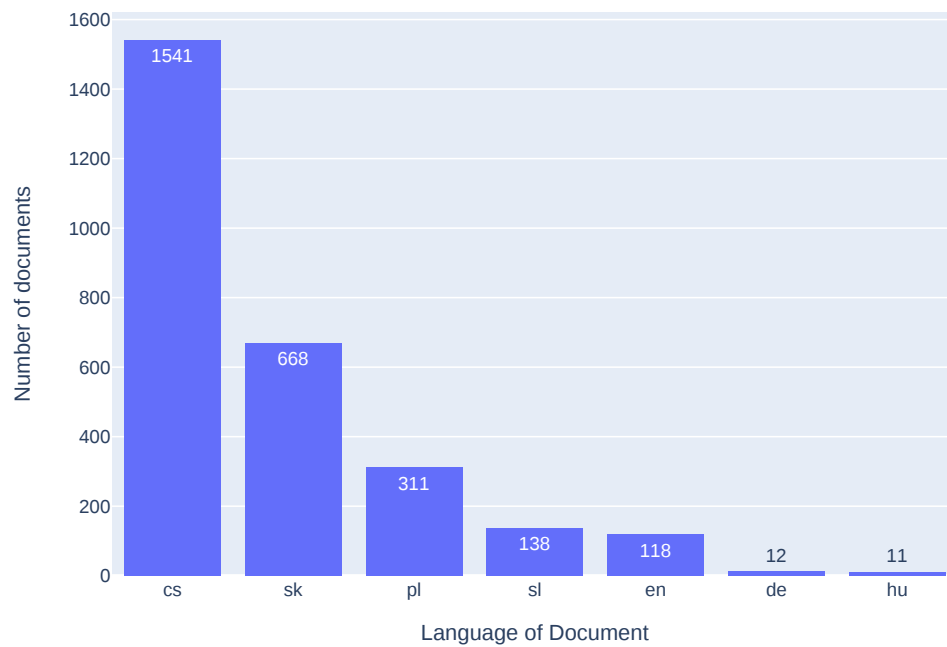


Fig. 3: Number of documents per language identified by language detection tool (visualizing only languages with more than 5 documents).



Table 1: Performance comparison of text recognition models on born-digital dataset

|              | Exact | Partial | FPS   |
|--------------|-------|---------|-------|
| Tesseract v5 | 0.90  | 0.90    | 3.35  |
| EasyOCR CRNN | 0.83  | 0.84    | 34.14 |
| Doctr CRNN   | 0.89  | 0.89    | 27.36 |
| Doctr MASTER | 0.98  | 0.99    | 0.46  |
| Doctr ViTSTR | 0.75  | 0.83    | 18.84 |

LayoutLM family, we fine-tuned two models pretrained on English scanned documents: LayoutLMv2 Base and LayoutLMv2 Large, and pretrained on a multilingual dataset: LayoutXLM Base.

We used the Tesseract OCR engine for all models to extract text information from the annotated scanned dataset.

In Table 2, we can see that pretrained multi-modal achieved higher scores than their equal-sized tex-only-based models. Furthermore, we see that both text-based and multi-modal models improved more by increasing the model size than by including multilingual pretraining since the best-performing text-only model is XLM RoBERTa Large, which is the strongest text-only model, and the overall best-performing model is LayoutLMv2 Large.

In Figure 4, we can see the confusion matrices of two best-performing multimodal models: LayoutXLM base and LayoutLMv2 large.

Table 2: Performance comparison of Text-based and LayoutLM models on separated evaluation datasets.

|                                     | F1-score     | Precision    | Recall       |
|-------------------------------------|--------------|--------------|--------------|
| <b>BERT Base Multilingual Cased</b> | 66.74        | 66.75        | 66.73        |
| <b>XLM RoBERTa Base</b>             | 72.80        | 72.61        | 73.00        |
| <b>RoBERTa Large</b>                | 78.25        | 77.52        | 79.00        |
| <b>XLM RoBERTa Large</b>            | 79.36        | 80.30        | 78.44        |
| <b>LayoutLM v2 Base</b>             | 77.83        | 75.99        | 79.76        |
| <b>LayoutLM v2 Large</b>            | <b>83.06</b> | <b>82.38</b> | <b>83.75</b> |
| <b>LayoutXLM Base</b>               | 79.40        | 78.75        | 80.06        |

## 5 Discussion and Future Work

In this paper, we presented end to end solution for information extraction in business documents. We offered solutions for both OCR and information extraction by text-only and multi-modal Transformers.

| Actual Category \ Predicted Category | DIC   | IBAN | ICD  | 0     | account_nu... | bank_code | const_symb... | contr_addre... | contr_name | due_date | invoice_date | invoice_nu... | qr_code | spec_symbol | total_amount | var_symbol |       |
|--------------------------------------|-------|------|------|-------|---------------|-----------|---------------|----------------|------------|----------|--------------|---------------|---------|-------------|--------------|------------|-------|
| DIC                                  | 82.84 | 0    | 1.25 | 15.06 | 0             | 0         | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0.83  |
| IBAN                                 | 0     | 97.2 | 0    | 2.23  | 0.27          | 0.27      | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| ICD                                  | 1.53  | 0    | 84   | 8.61  | 0.3           | 0         | 0.61          | 0.61           | 1.84       | 0        | 0            | 0.3           | 0       | 0           | 0            | 0          | 2.15  |
| 0                                    | 0.1   | 0.05 | 0.06 | 98.07 | 0.03          | 0         | 0.01          | 0.5            | 0.6        | 0.07     | 0.1          | 0.04          | 0.03    | 0           | 0.24         | 0.02       | 0     |
| account_number                       | 4.22  | 0    | 0    | 22.53 | 72.53         | 0.7       | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| bank_code                            | 0     | 0    | 0    | 15.78 | 7.36          | 76.84     | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| const_symbol                         | 0     | 0    | 0    | 12.96 | 0             | 0         | 87.03         | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| contr_address                        | 0     | 0    | 0.05 | 10.19 | 0             | 0         | 88.04         | 1.69           | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| contr_name                           | 0     | 0    | 0    | 14.06 | 0             | 0         | 1.94          | 83.98          | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| due_date                             | 0     | 0    | 0    | 11.34 | 0             | 0         | 0             | 0              | 87.73      | 0.92     | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| invoice_date                         | 0     | 0    | 0.24 | 15.13 | 0             | 0         | 0.24          | 0.49           | 2.43       | 81.38    | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| invoice_number                       | 0     | 0    | 0    | 12.2  | 0             | 0.33      | 0             | 0              | 0          | 0        | 82.71        | 0             | 0       | 0           | 0            | 0          | 4.74  |
| qr_code                              | 0     | 0    | 0    | 7.2   | 0             | 0         | 0             | 0              | 4          | 0        | 0            | 24            | 0       | 0           | 0            | 0          | 0     |
| spec_symbol                          | 0     | 0    | 0    | 100   | 0             | 0         | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| total_amount                         | 0     | 0    | 0.09 | 11.63 | 0             | 0.09      | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 88.17 |
| var_symbol                           | 0     | 0.76 | 0    | 5.38  | 0             | 0         | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 90    |

| Actual Category \ Predicted Category | DIC   | IBAN  | ICD   | 0     | account_nu... | bank_code | const_symb... | contr_addre... | contr_name | due_date | invoice_date | invoice_nu... | qr_code | spec_symbol | total_amount | var_symbol |       |
|--------------------------------------|-------|-------|-------|-------|---------------|-----------|---------------|----------------|------------|----------|--------------|---------------|---------|-------------|--------------|------------|-------|
| DIC                                  | 85.65 | 0.43  | 2.6   | 10.43 | 0             | 0         | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0.86  |
| IBAN                                 | 0     | 97.37 | 0     | 2.62  | 0             | 0         | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| ICD                                  | 1.47  | 0.73  | 83.08 | 9.55  | 0             | 0         | 0             | 0              | 4.77       | 0.36     | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| 0                                    | 0.07  | 0.03  | 0.04  | 98.38 | 0.03          | 0         | 0.02          | 0.34           | 0.54       | 0.04     | 0.12         | 0.08          | 0       | 0           | 0.25         | 0.01       | 0     |
| account_number                       | 0.4   | 0.4   | 0     | 21.37 | 77.01         | 0.8       | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| bank_code                            | 0     | 0     | 0     | 29.21 | 6.74          | 64.04     | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| const_symbol                         | 0     | 0     | 0     | 3.77  | 0             | 0         | 96.22         | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| contr_address                        | 0     | 0     | 0.18  | 10.8  | 0             | 0         | 87.42         | 1.58           | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| contr_name                           | 0     | 0     | 0     | 14.49 | 0             | 0         | 2.14          | 83.35          | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| due_date                             | 0     | 0     | 0     | 9.03  | 0             | 0         | 0             | 0              | 0          | 90.32    | 0.64         | 0             | 0       | 0           | 0            | 0          | 0     |
| invoice_date                         | 0.27  | 0     | 0.27  | 11.5  | 0             | 0         | 0.27          | 0.54           | 0.27       | 86.57    | 0            | 0             | 0       | 0           | 0            | 0          | 0.27  |
| invoice_number                       | 0     | 0     | 0     | 10.8  | 0             | 0         | 0             | 0              | 0          | 0        | 88.15        | 0             | 0       | 0           | 0            | 0          | 1.04  |
| qr_code                              | 0     | 0     | 0     | 85.71 | 0             | 0         | 0             | 0              | 0          | 0        | 0            | 14.28         | 0       | 0           | 0            | 0          | 0     |
| spec_symbol                          | 0     | 0     | 0     | 100   | 0             | 0         | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 0     |
| total_amount                         | 0     | 0     | 0     | 13.52 | 0             | 0.14      | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 0            | 0          | 86.32 |
| var_symbol                           | 0     | 0     | 0     | 2.34  | 0             | 0         | 0             | 0              | 0          | 0        | 0            | 0             | 0       | 0           | 1.56         | 0          | 96.09 |

Fig. 4: Confusion matrices of finetuned LayoutXLM base (left) and LayoutLMv2 large (right) models.

In the OCR sections of the paper, we proved that EasyOCR and Doctr framework are sufficient for document understanding. However, to create novel and well-behaving products, custom training is necessary. Both have the common intersection of models; however, models are not compatible with each other and prevent any sensible comparison. Training scripts for EasyOCR need to be more documented. Doctr framework lacks parametrization. The training can be improved by hyper-parameter search, as Transformers and CRNN have a different sensibility to learning rate and the number of epochs. Our paper does not evaluate the text detection models, which can cause a bottleneck in the commercial use of the system.

In the NER sections of the paper, we offered an overview of available multi-modal Transformer models and, more in detail described the family of the LayoutLM models. In the Section 4, the LayoutLM model with equal size achieved significantly better results than their equally-sized text-only counterparts. We also show that on presented dataset size of the model improved results by a higher margin than introducing multilingual pretraining. This behavior can be explained by the types of extracting entities, mainly composed of information written in digits (number codes, dates, times, sum), which are language-independent.

In future work, we propose experimenting with the LayoutLMv3 model as we revealed that models pretrained only on English could outperform multilingual-based models. Furthermore, we plan to research the impact of the used OCR engine on the performance of the NER model.

**Acknowledgements** We acknowledge the support of grant Intelligent Back Office, project number CZ.01.1.02/0.0/0.0/21\_374/0026711.

## References

1. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: DocFormer: End-to-End Transformer for Document Understanding. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 973–983 (2021). <https://doi.org/10.1109/ICCV48922.2021.00103>
2. Atienza, R.: Vision transformer for fast and efficient scene text recognition. In: International Conference on Document Analysis and Recognition. pp. 319–334. Springer (2021)
3. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character Region Awareness for Text Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. CORR (2020), <https://arxiv.org/abs/2010.11929v2>
5. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091. MM '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3503161.3548112>
6. Jaided AI: Easyocr. <https://github.com/JaidedAI/EasyOCR> (2020)
7. Lee, C.Y., Li, C.L., Dozat, T., Perot, V., Su, G., Hua, N., Ainslie, J., Wang, R., Fujii, Y., Pfister, T.: FormNet: Structural encoding beyond sequential modeling in form document information extraction. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3735–3754. ACL, Dublin, Ireland (May 2022). <https://doi.org/10.18653/v1/2022.acl-long.260>, <https://aclanthology.org/2022.acl-long.260>
8. Li, P., Gu, J., Kuen, J., Morariu, V.I., Zhao, H., Jain, R., Manjunatha, V., Liu, H.: SelfDoc: Self-Supervised Document Representation Learning (2021). <https://doi.org/10.48550/ARXIV.2106.03331>
9. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2022). <https://doi.org/10.1109/TPAMI.2022.3155612>
10. Lu, N., Yu, W., Qi, X., Chen, Y., Gong, P., Xiao, R., Bai, X.: Master: Multi-aspect non-local network for scene text recognition. Pattern Recognition **117**, 107980 (2021)
11. Mindee: docTR: Document Text Recognition. <https://github.com/mindee/doctr> (2021)
12. Shi, B., Bai, X., Yao, C.: An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(11), 2298–2304 (2017). <https://doi.org/10.1109/TPAMI.2016.2646371>
13. Smith, R.: An overview of the Tesseract OCR engine. In: Ninth international conference on document analysis and recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
14. Soboroff, I.: Complex Document Information Processing (CDIP) dataset (2022). <https://doi.org/10.18434/mds2-2531>
15. Wang, J., Jin, L., Ding, K.: LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding (2022). <https://doi.org/10.48550/ARXIV.2202.13669>

16. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., Zhou, L.: LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In: Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2579–2591. ACL, Online (Aug 2021). <https://doi.org/10.18653/v1/2021.acl-long.201>, <https://aclanthology.org/2021.acl-long.201>
17. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: Pre-Training of Text and Layout for Document Image Understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200. KDD '20, Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3394486.3403172>
18. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florêncio, D., Zhang, C., Wei, F.: LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. CoRR [abs/2104.08836](https://arxiv.org/abs/2104.08836) (2021), <https://arxiv.org/abs/2104.08836>