MUNI
FI

Information Extraction from Business Documents

Case Study

M. Geletka, M. Bankovič, D. Meluš, Š. Ščavnická, M. Štefánik, P. Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

December 9th, 2022

Problem definition

- **Problem**: Read and understand business documents
- **Goal**: Automatization of payments from scanned invoices
- **Possible solutions**:
  - OCR -> Text only NER
  - OCR -> Token classification (based on text and position)
  - OCR -> MultiModal NER (based om position, image and text)

Deep Learning OCR pipeline

- Pre-processing (Minimal)
- Text Detection
- Text Recognition
- Post-processing

Deep Learning OCR

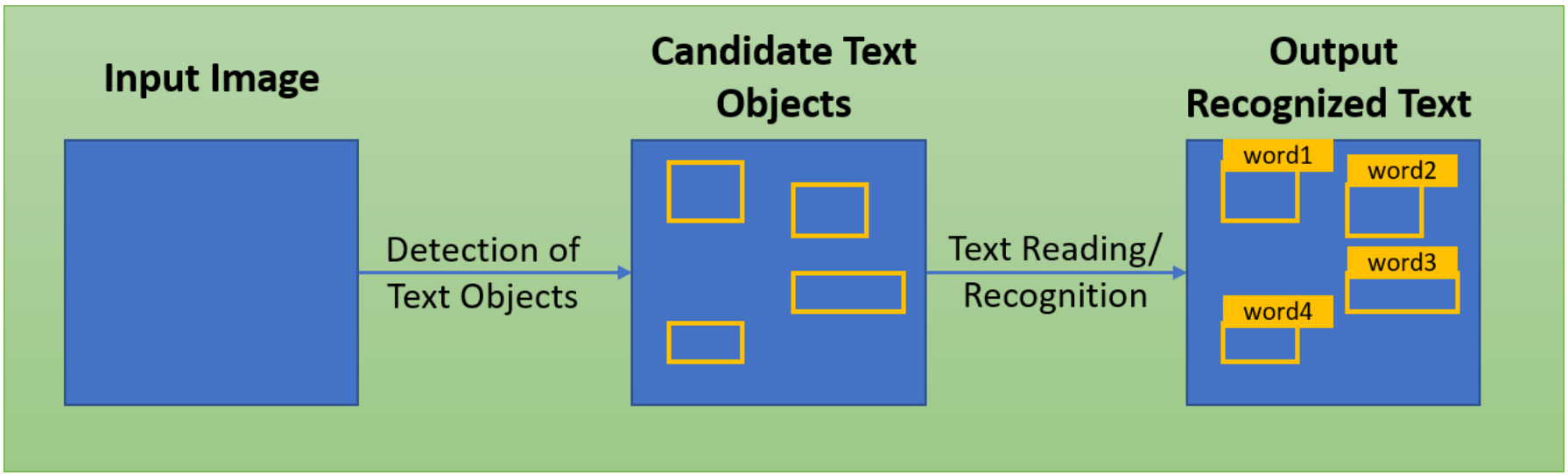


Figure: Deep Learning OCR pipeline

OCR text recognition comparison

- Tesseract Version 5.0
  - CRNN (CNN + LSTM)
- EasyOCR
  - CRNN
- Doctr
  - CRNN, MASTER and ViTSTR

Born Digital Dataset

- we selected digitally-born documents, PDF files
- we extracted bounding boxes and letter via available python tools
- we filtered bad extraction, hidden symbols and got around 600k bounding box and text pairs (30k unique words)

OCR Results

Table: 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

Multi-modal Information Extraction

- family of models, which use additional modalities compare to text only models
- modalities used:
  - text – WordPiece tokens
  - position of text – 2D position in the Document
  - image – vectorized representations of the image

LayoutLM family

- family of multi-modal models from Microsoft:
- English models
  - LayoutLMv1 [8]
  - LayoutLMv2 [7]
  - LayoutLMv3 [2]
- Multilingual models
  - LayoutXLM [9]

Layout(X)LM training

- Dataset
  - English
    - IIT-CDIP Test Collection 1.0 – 11 M scanned documents (6 M unique) [5]
  - Multilingual
    - born-digital documents crawled from web
    - including 53 languages
- Pretraining tasks
  - Masked Visual-Language Model
  - Multi-label Document Classification
  - Text-Image Alignment
  - Text-Image Matching

Layout (X)LM architecture

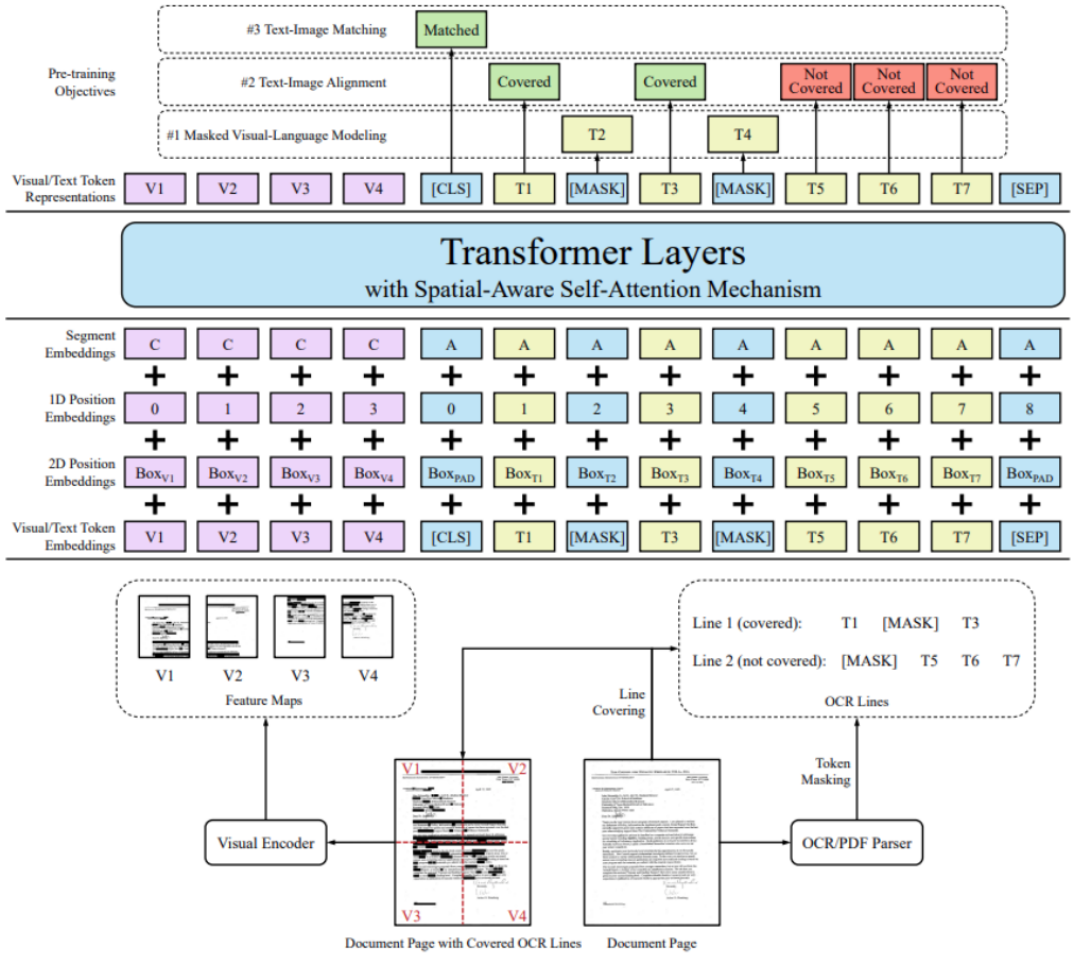


Figure: Example of one input data sample

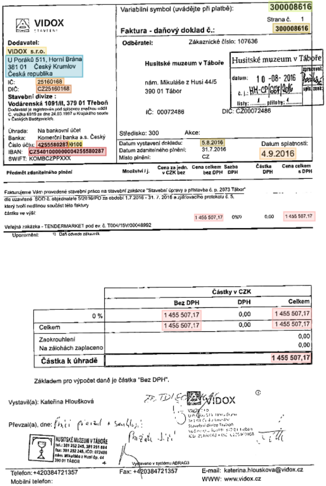
Other Related Work

- MultiLingual
  - **LiLT** [6] - LayoutLM like model with independant text model
- English
  - **FormNet** [3] - models relationship between tokens by graph convolution
  - **DocFormer** [1] -similar model to LayoutLM family with different implemetnation of multimodal attention
  - **SelfDoc** [4] - uses larger semantic components instead of WordPiece tokens

Dataset description

- we collected of set of documents, as no Czech documents dataset of a sufficient volume or quality is available
- dataset obtained by querying server `uloz.to` with keywords “faktura”, “smlouva” or “doklad”
- result collection after manually cleaning contains 6,849 invoice images
- dataset annotation
  - selecting a bounding box
  - assigning a category

Dataset sample



VIDOX

Dodavatel:
VIDOX s.r.o.
U Poráků 511, Horní Brána
381 01 Český Krumlov
Česká republika
IČ: 25160168
DIČ: CZ25160168
Stavební divize :
Vodárenská 1091/II, 379 01 Třeboň

Úhrada: Na bankovní účet
Banka: Komerční banka a.s. Český
Číslo účtu: 4255580287/0100
IBAN: CZ5401000000004255580287
SWIFT: KOMBCZPPXXX

Variabilní symbol (uvádějte při platbě): 300008616
Strana č. 1
Faktura - daňový doklad č.: 300008616

Odběratel: Zákaznické číslo: 107636

Husitské muzeum v Táboře
nám. Mikuláše z Husi 44/5
390 01 Tábor

IČ: 00072486 DIČ: CZ00072486

Středisko: 300 Akce:
Datum vystavení dokladu: 5.8.2016
Datum zdanitelného plnění: 31.7.2016
Místo plnění: CZ
Datum splatnosti: 4.9.2016

Předmět zdanitelného plnění	Množství j.	Cena za jedn. v CZK bez	Cena celkem bez DPH	Sazba DPH	Částka DPH	Cena celkem s DPH
Fakturujeme Vám provedené stavební práce na stavební zakázce "Stavební úpravy a přístavba č. p. 2073 Tábor" dle uzavřené SOD č. objednatele 5/2016/PO za období 1.7.2016 - 31. 7. 2016 a zjišťovacího protokolu č. 5, který tvoří nedílnou součást této faktury částku ve výši:			1 455 507,17	0%1)	0,00	1 455 507,17

Veřejná zakázka - TENDERMARKET pod ev. č. T004/15V/00048992

Upozornění: 1) Daň odvede zákazník

	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 zaplaceno			0,00
Částka k úhradě			1 455 507,17

Základem pro výpočet daně je částka "Bez DPH".

Vystavil(a): Kateřina Hloušková

Převzal(a), dne:

Telefon: +420384721357
Mobilní telefon:
Fax: +420384721357
E-mail: katerina.hlouskova@vidox.cz
WWW: www.vidox.cz

Figure: Example of one input data sample

Dataset languages and entity types

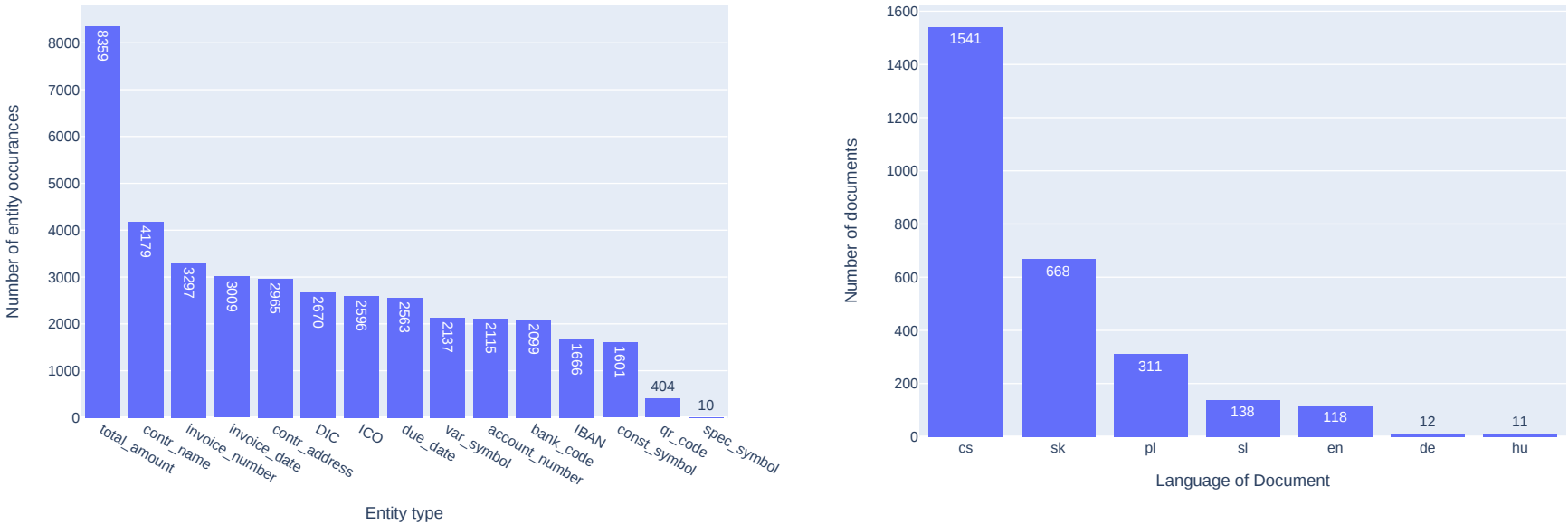


Figure: Histograms of entity types (left) and languages of scrapped documents (right).

Results

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

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

Conclusion & Future work

- Future work OCR
  - train EasyOCR model
  - data augmentation to imitate bad scans while training with born-digital documents
  - use language model as post-processing
  - End2End text detection and text recognition evaluation - number of found annotated entities
- Future work NER
  - implement post-processing merging for output bounding boxes
  - train LiLT model and compare to LayoutLm results
  - compare models trained with different OCR engines
- Common Future work
  - end2end evaluation – such as number of successful automatic payment transactions

Bibliography I

[1] Srikar Appalaraju et al. “DocFormer: End-to-End Transformer for Document Understanding”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 973–983. DOI: 10.1109/ICCV48922.2021.00103.

[2] Yupan Huang et al. “LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. MM ’22. Lisboa, Portugal: Association for Computing Machinery, 2022, pp. 4083–4091. ISBN: 9781450392037. DOI: 10.1145/3503161.3548112.

Bibliography II

[3] Chen-Yu Lee et al. “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)*. Dublin, Ireland: ACL, May 2022, pp. 3735–3754. DOI: 10.18653/v1/2022.acl-long.260. URL: https://aclanthology.org/2022.acl-long.260.

[4] Peizhao Li et al. *SelfDoc: Self-Supervised Document Representation Learning*. 2021. DOI: 10.48550/ARXIV.2106.03331.

[5] Ian Soboroff. *Complex Document Information Processing (CDIP) dataset*. 2022. DOI: 10.18434/mds2-2531. (Visited on 11/10/2022).

Bibliography III

[6] Jiapeng Wang, Lianwen Jin, and Kai Ding. *LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding*. 2022. DOI: 10.48550/ARXIV.2202.13669.

[7] Yang Xu et al. “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)*. Online: ACL, Aug. 2021, pp. 2579–2591. DOI: 10.18653/v1/2021.acl-long.201. URL: https://aclanthology.org/2021.acl-long.201.

Bibliography IV

[8] Yiheng Xu et al. “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*. KDD ’20. Virtual Event, CA, USA: Association for Computing Machinery, 2020, pp. 1192–1200. ISBN: 9781450379984. DOI: 10.1145/3394486.3403172.

[9] Yiheng Xu et al. “LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding”. In: *CoRR* abs/2104.08836 (2021). arXiv: 2104.08836. URL: https://arxiv.org/abs/2104.08836.

Thank you for your attention!

MUNI
FACULTY
OF INFORMATICS