Training LayoutLM from scratch for efficient Named-Entity Recognition in the Insurance domain

Benno Uthayasooriyar[1,2], Antoine Ly[1], Franck Vermet[2], Caio Corro[3]

[1]AI & Advanced Analytics, SCOR [2]Univ Brest, CNRS, UMR 6205, LMBA [3]IRISA, Inria, CNRS, Université de Rennes

Contributions

1. We demonstrate that domain specific pre-training improves performances and reduces results variances on our target task, with less data
2. We achieve comparable results with a smaller, faster model, better suited for real-time processing of large documents volumes
3. We introduce **PAYSLIPS**, a novel NER dataset for financial documents

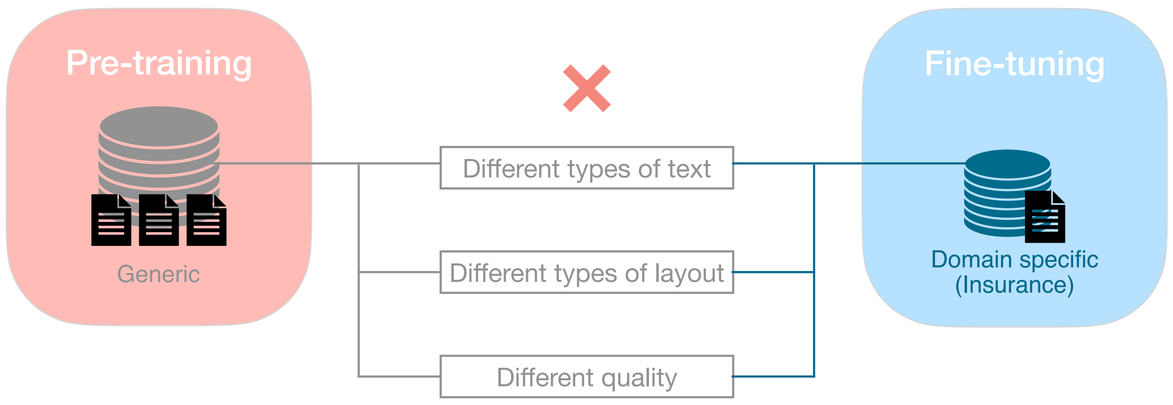
Named-Entity Recognition

We are interested in Named-Entity Recognition (NER) to extract information from insurance documents.

Donald Trump PER is the president of the USA LOC.

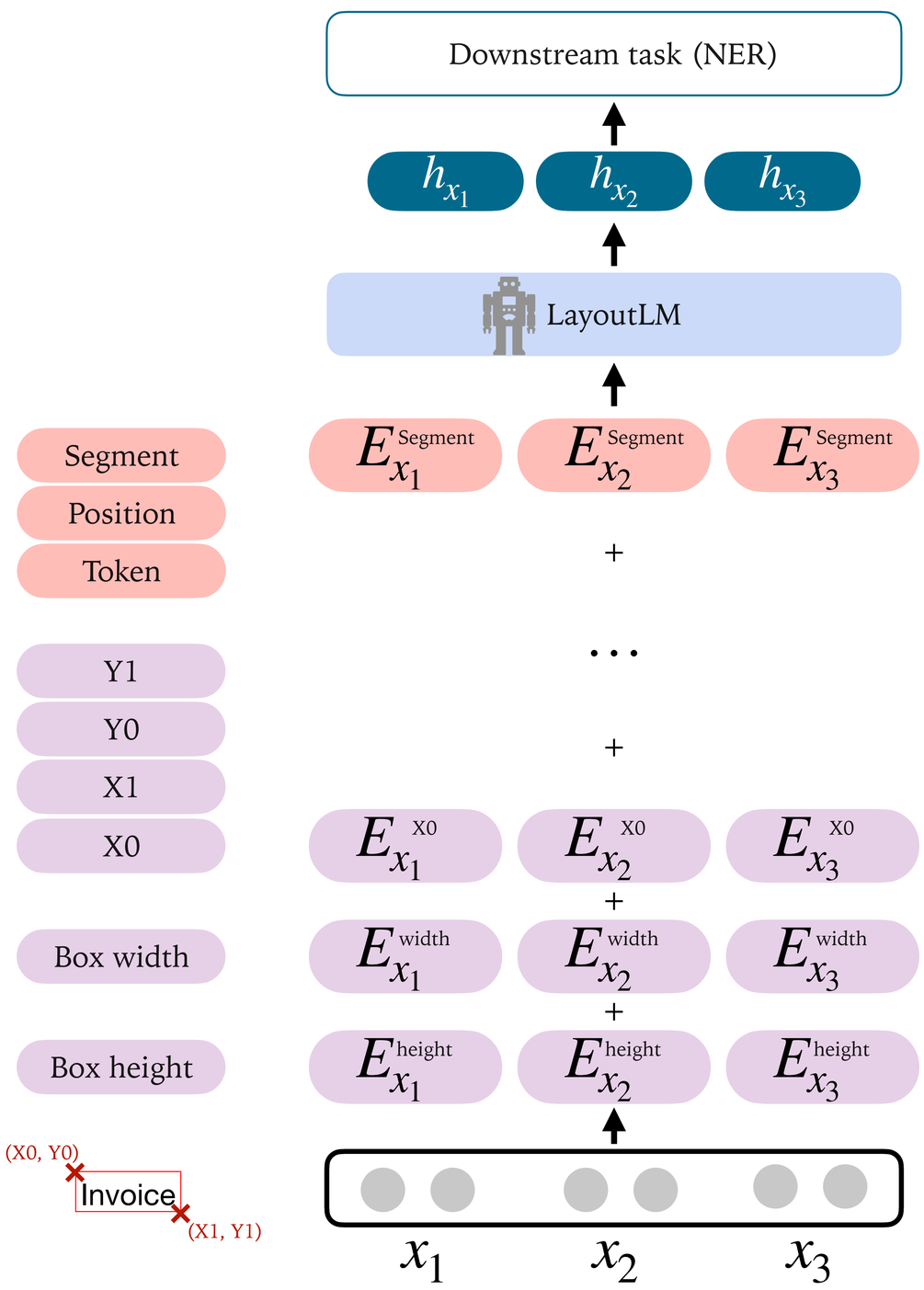
Problem

Layout information is crucial for insurance documents processing. However, layout-aware pre-trained language models can struggle on specialized domains due to mismatches between training data and downstream task.



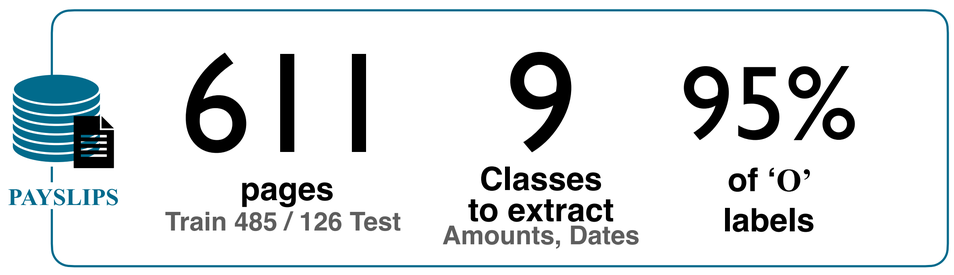
Document Understanding Models

We use **LayoutLM**, an extension of BERT that adds layout information in the input embedding, along a classifier to perform the NER task.



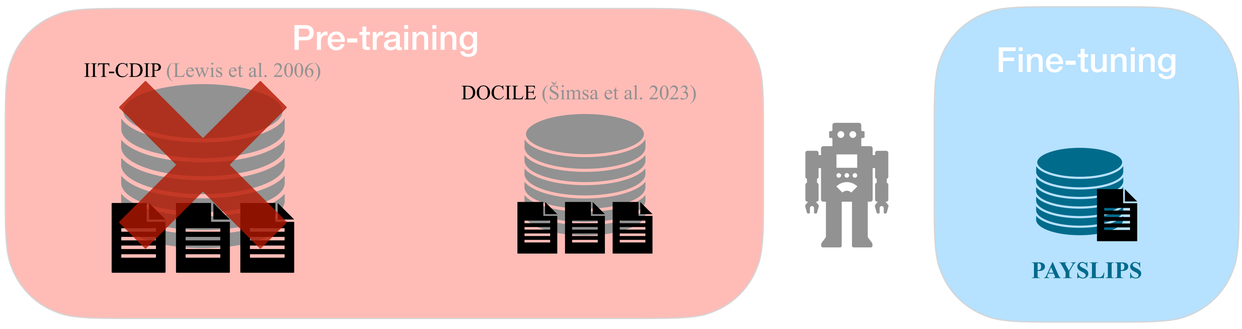
The PAYSLIPS Dataset

We release a novel dataset of anonymous pay statements from an insurance context, with layout and text annotations for NER.



In-domain pre-training

We performed pre-training on **DOCILE**, a dataset of ~900k invoices, instead of the traditionally used **IIT-CDIP** dataset made of 11M tobacco industry documents from the 90s.



Experimental Results

Results show improved performances with pre-training on DOCILE, with **less data** that are visually and semantically closer to the downstream task.

Model	F1 DOCILE labeled	F1 PAYSLIPS
Pre-training on IIT-CDIP		
LAYOUTLM$_{\text{BASE}}$	**58.35 ± 1.63**	62.31 ± 5.13
Pre-training on DOCILE		
LAYOUTLM$_{\text{BASE}}$	58.30 ± 1.52	**64.74 ± 2.92**
LAYOUTLM$_{\text{6 layers}}$	57.38 ± 1.38	61.80 ± 3.12
LAYOUTLM$_{\text{2 layers}}$	53.89 ± 1.03	54.61 ± 3.71
LAYOUTLM$_{\text{1 layer}}$	51.12 ± 1.53	45.08 ± 3.31

To ensure resistance to the high flow of documents of a commercial use, we experimented with reduced number of self-attentive layers.
The 6 layers model maintains results comparable to the off-the-shelf model, while being twice as fast.

Model	Inference Time (ms)
LAYOUTLM$_{\text{BASE}}$	12.10
LAYOUTLM$_{\text{6 layers}}$	6.15
LAYOUTLM$_{\text{2 layers}}$	2.42
LAYOUTLM$_{\text{1 layers}}$	1.73

Thanks to these learnings, future work will involve a novel model architecture with improved layout aware self-attention.

arXiv GitHub

Acknowledgments

This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011015001)