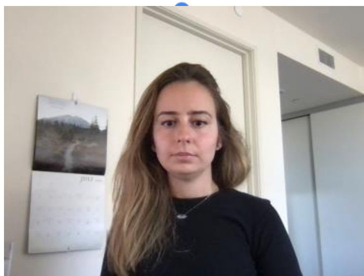
Google Research

Data-Efficient Information Extraction from Form-Like Documents

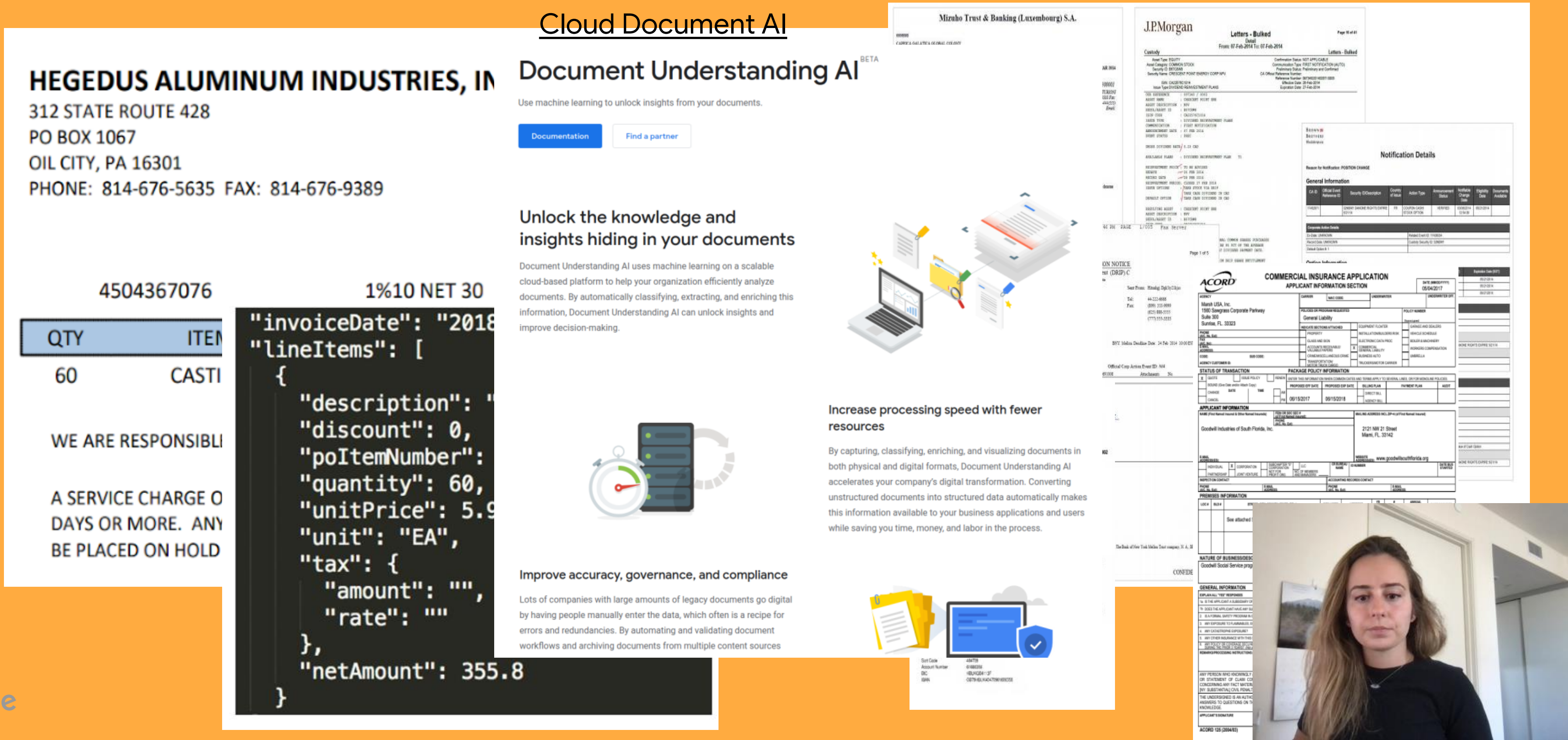
Beliz Gunel, Navneet Potti, Sandeep Tata, James B. Wendt, Marc Najork, Jing Xie

KDD-DI Workshop 2021, Machine Learning Session



Automating information extraction from **form-like** documents *at scale* can have a huge impact on business workflows.

Cloud Document AI



Holistic understanding of textual segments & visual cues within a document is non-trivial.

LayoutLM: Pre-training of Text and Layout for Document Image Understanding

Yiheng Xu[*]
charlesyihengxu@gmail.com
Harbin Institute of Technology

Minghao Li[*]
liminghao1630@buaa.edu.cn
Beihang University

Lei Cui
lecu@microsoft.com
Microsoft Research Asia

Shaohan Huang
shaohanh@microsoft.com
Microsoft Research Asia

Furu Wei
fuwei@microsoft.com
Microsoft Research Asia

Ming Zhou
mingzhou@microsoft.com
Microsoft Research Asia

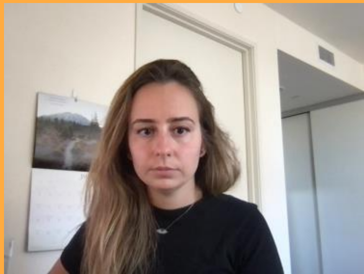
BERTgrid: Contextualized Embedding for 2D Document Representation and Understanding

Timo I. Denk[*]
SAP SE
Machine Learning R&D, Berlin Germany
mail@timodenk.com

Christian Reisswig[*]
SAP SE
Machine Learning R&D, Berlin Germany
christian.reisswig@sap.com

Representation Learning for Information Extraction from Form-like Documents

Bodhisattwa Prasad Majumder[†♣] Navneet Potti[♠] Sandeep Tata[♠]
James B. Wendt[♠] Qi Zhao[♠] Marc Najork[♠]
[♣]Department of Computer Science and Engineering, UC San Diego
bmajumde@eng.ucsd.edu
[♠]Google Research, Mountain View
{navsan, tata, jwendt, zhaqi, najork}@google.com



Main cost is data acquisition and labeling for every new language or every new document type.

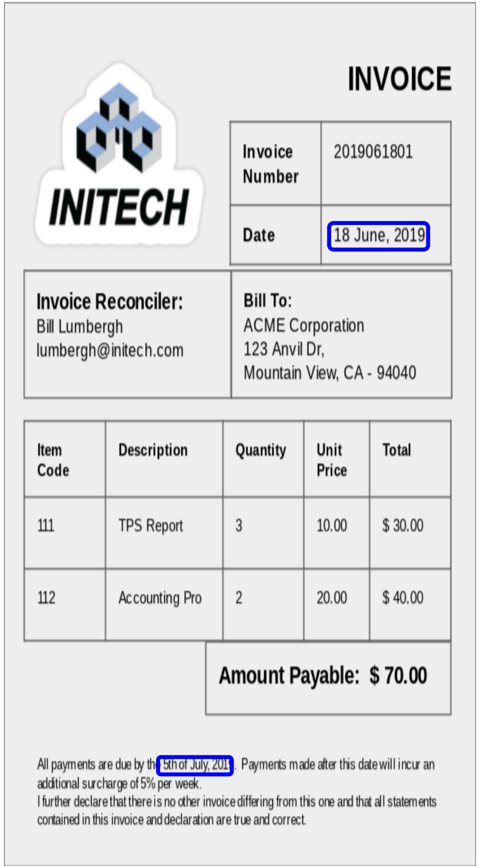
Previous approaches are promising, but training/pre-training part of their pipelines are
(1) compute-intensive
(2) data-intensive
(3) re-done from scratch for competitive performance for every new language/doc type

If we can get to same extraction performance with 10x less data, we effectively cut the cost of developing new extraction models by 10x. Hence, this paper focuses on:

(1) data-efficiency
(2) ability to generalize across different document types and languages

We build on Glean Extraction Pipeline

Google Research


INITECH

INVOICE

Invoice Number	2019061801
Date	18 June, 2019

Invoice Reconciler:
Bill Lumbergh
lumbergh@initech.com

Bill To:
ACME Corporation
123 Anvil Dr,
Mountain View, CA - 94040

Item Code	Description	Quantity	Unit Price	Total
111	TPS Report	3	10.00	$ 30.00
112	Accounting Pro	2	20.00	$ 40.00

Amount Payable: $ 70.00

All payments are due by the 5th of July, 2019. Payments made after this date will incur an additional surcharge of 5% per week.
I further declare that there is no other invoice differing from this one and that all statements contained in this invoice and declaration are true and correct.

1. Generate Candidates Based on Field Type

18 June, 2019

5 July, 2019

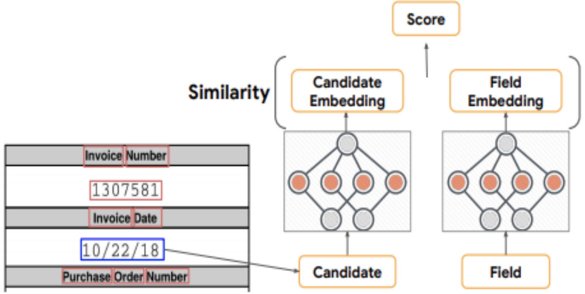
2. Score Each Candidate Based on Its Neighborhood

18 June, 2019 0.3

5 July, 2019 0.9

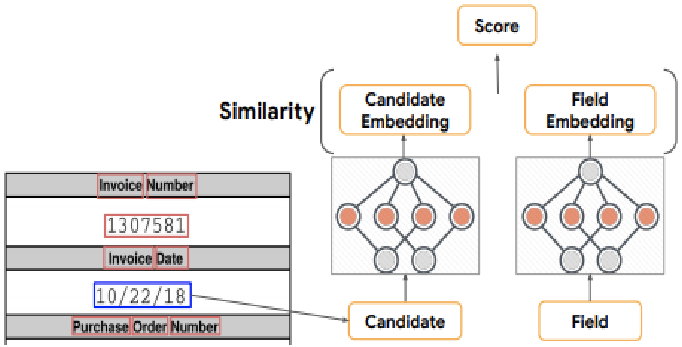
3. Assign Highest-Scoring Candidate as Extraction Result

5 July, 2019



Main Hypothesis: Form-like documents share a visual design language, hence we can effectively transfer knowledge across considerably different domains.

Our Proposal: Multi-domain Transfer Learning

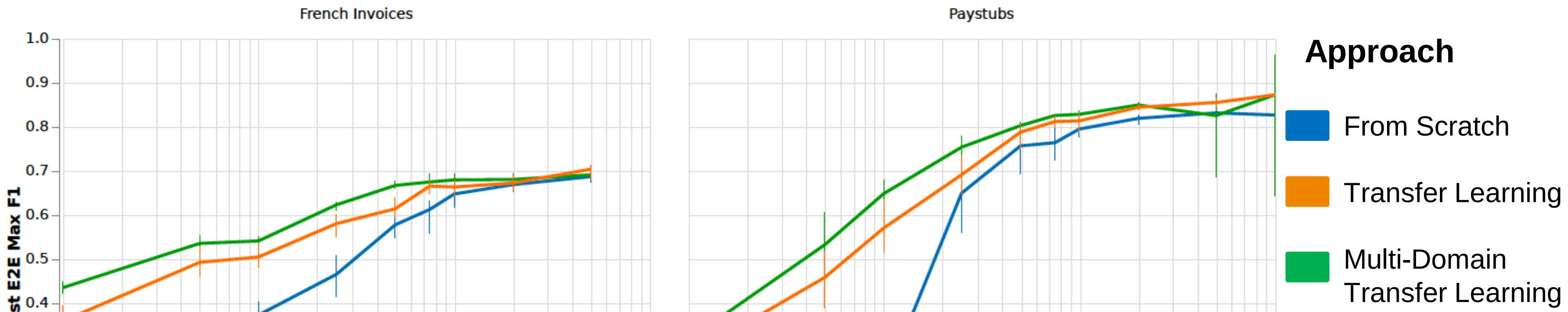


<table>
<tr><th>Approach</th><th>Initial Training Stage</th><th>Fine-tuning Stage</th></tr>
<tr><td>From Scratch</td><td>-</td><td rowspan="3">Target domain only</td></tr>
<tr><td>Transfer Learning</td><td>Source domain only</td></tr>
<tr><td>Multi-domain Transfer Learning</td><td>Source & target domains</td></tr>
</table>

<u>Initial Training Stage:</u> Learn a candidate encoder that learns to represent domain-agnostic spatial relationships between candidate and its neighbors.
<u>Fine-tuning Stage:</u> Fine-tune learned candidate encoder and field embeddings on the domain of interest.

Use a common vocabulary across source & target domains.

Results



We improve on the training from scratch baseline by up to <u>35 F1 points</u>, and on the simple transfer learning baseline by up to <u>8 F1 points</u> for the 50 labeled document case while generalizing to a new document type; training from scratch baseline by up to <u>23 F1 points</u>, and on the simple transfer learning baseline by up to <u>7 F1 points</u> for the 10 labeled document case while generalizing to a new language.

Model training takes 45 mins on a single GPU + approach is currently in production use.

Future Work

- Data efficiency will be increasingly more critical as information extraction systems will need to perform well across ***more document types, more languages, and potentially on private customer data.***
- Next big step: Decreasing the labeled document need from ~1K to ~100 for each (n+1)th document type or language we would like to generalize to.

Thanks for listening!

I am broadly interested in **representation learning** and its applications for healthcare, natural language processing, and **building data-efficient machine learning methods that are robust to distribution drifts**.

Beliz Gunel

PhD Candidate @ Stanford

I will be graduating late 2022!

http://web.stanford.edu/~bgunel/