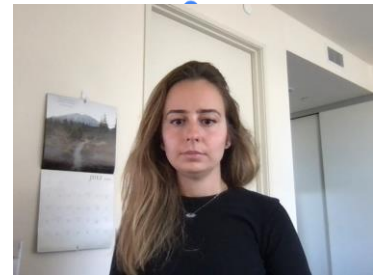




Data-Efficient Information Extraction from Form-Like Documents

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

Document Understanding AI

Use machine learning to unlock insights from your documents.

Documentation Find a partner

Unlock the knowledge and insights hiding in your documents

Document Understanding AI uses machine learning on a scalable cloud-based platform to help your organization efficiently analyze documents. By automatically classifying, extracting, and enriching this information, Document Understanding AI can unlock insights and improve decision-making.



Increase processing speed with fewer resources

By capturing, classifying, enriching, and visualizing documents in both physical and digital formats, Document Understanding AI accelerates your company's digital transformation. Converting unstructured documents into structured data automatically makes this information available to your business applications and users while saving you time, money, and labor in the process.

Improve accuracy, governance, and compliance

Lots of companies with large amounts of legacy documents go digital by having people manually enter the data, which often is a recipe for errors and redundancies. By automating and validating document workflows and archiving documents from multiple content sources



HEGEDUS ALUMINUM INDUSTRIES, INC
 312 STATE ROUTE 428
 PO BOX 1067
 OIL CITY, PA 16301
 PHONE: 814-676-5635 FAX: 814-676-9389

4504367076

1%10 NET 30

QTY	ITEM
60	CASTI

WE ARE RESPONSIBL

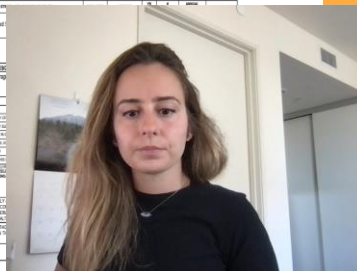
A SERVICE CHARGE O
 DAYS OR MORE. ANY
 BE PLACED ON HOLD

```

    "invoiceDate": "2018
    "lineItems": [
      {
        "description": "
        "discount": 0,
        "poItemNumber":
        "quantity": 60,
        "unitPrice": 5.9
        "unit": "EA",
        "tax": {
          "amount": "",
          "rate": ""
        },
        "netAmount": 355.8
      }
    ]
  
```

The collage includes:

- Mimbo Trust & Banking (Luxembourg) S.A. document
- JPMorgan Letters - Bullied document
- Notification Details form
- COMMERCIAL INSURANCE APPLICATION form
- ACCORD logo
- South USA, Inc. document
- Various other forms and documents.



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

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

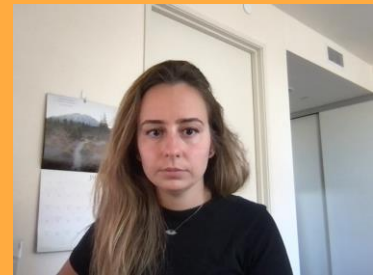
[♠]Google Research, Mountain View

{navsan, tata, jwendt, zhaqi,
najork}@google.com

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



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

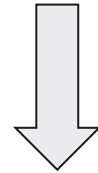
					INVOICE				
Invoice Number		2019061801							
Date		18 June, 2019							
Invoice Reconciler: Bill Lumbergh lumbergh@initech.com					Bill To: ACME Corporation 123 Anvil Dr, Mountain View, CA - 94040				
Item Code	Description	Quantity	Unit Price	Total					
111	TPS Report	3	10.00	\$ 30.00					
112	Accounting Pro	2	20.00	\$ 40.00					
Amount Payable: \$ 70.00									
All payments are due by the <u>18th July 2019</u> . 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

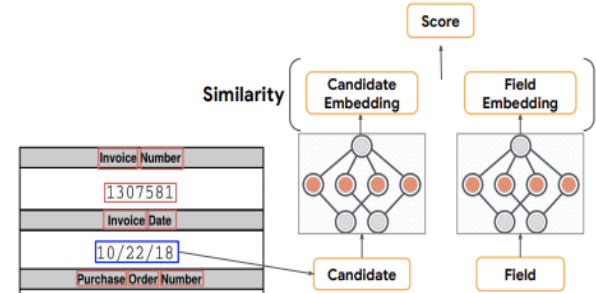


- 18 June, 2019 **0.3**
- 5 July, 2019 **0.9**

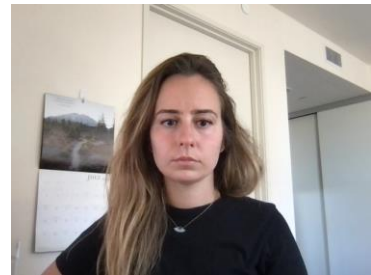
3. Assign Highest-Scoring Candidate as Extraction Result



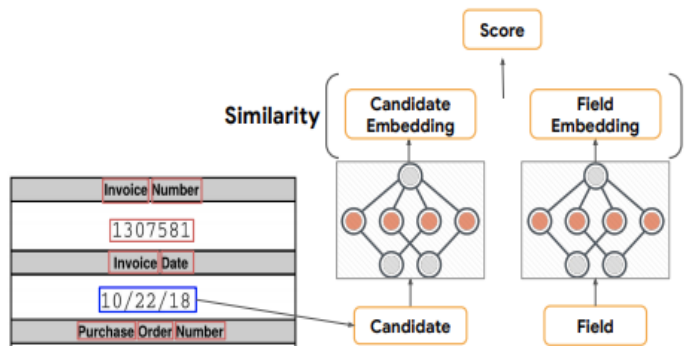
- 5 July, 2019



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



Approach	Initial Training Stage	Fine-tuning Stage
From Scratch	-	Target domain only
Transfer Learning	Source domain only	
Multi-domain Transfer Learning	Source & target domains	

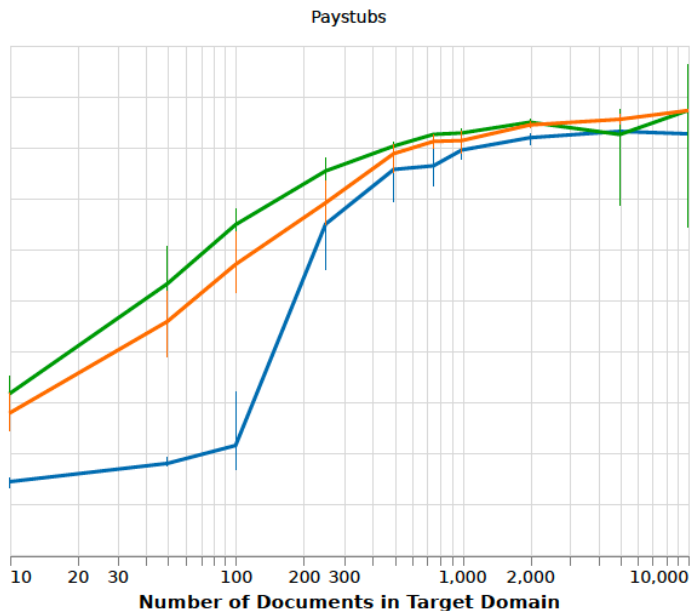
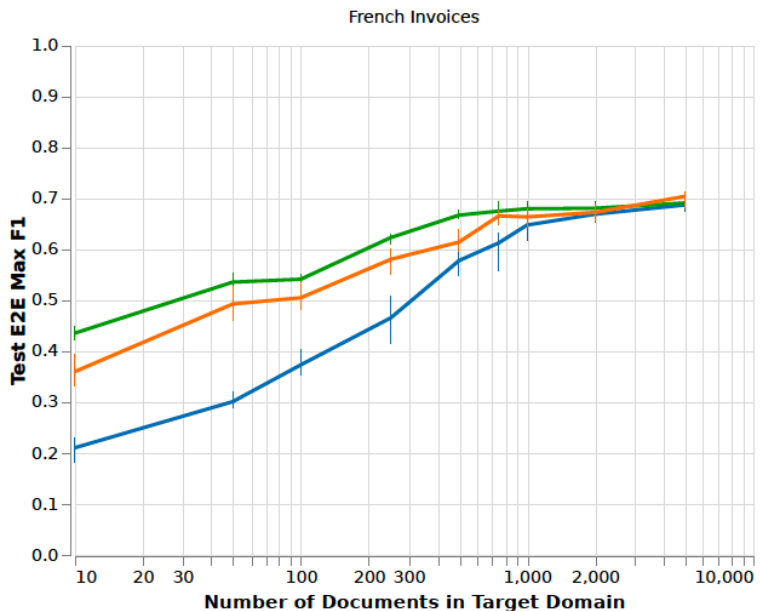
Initial Training Stage: Learn a candidate encoder that learns to represent domain-agnostic spatial relationships between candidate and its neighbors.

Fine-tuning Stage: Fine-tune learned candidate encoder and field embeddings on the domain of interest.

Use a common vocabulary across source & target domains.

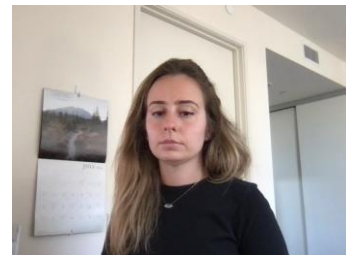


Results



Approach

- From Scratch
- Transfer Learning
- Multi-Domain Transfer Learning



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

