Document Intelligence Metrics

J. DeGange, S. Gupta, Z. Han, K. Wilkosz and A. Karwan
jonathan.degange@ey.com, swapnil.gupta@nyu.edu, zhuoyu.han@walmart.com,
krzysztof.wilkosz@allegro.pl, adam.karwan@gds.ey.com

Abstract

The processing of Visually-Rich Documents (VRDs) is highly important in information extraction tasks associated with Document Intelligence. We introduce **DI Metrics**, a Python library devoted to VRD model evaluation comprising text-based, geometric-based and hierarchical metrics for information extraction tasks. We apply **DI Metrics** to evaluate information extraction performance using publicly available *CORD* dataset, comparing performance of three SOTA models and one industry model.
The open-source library is available on GitHub[a].

[a] https://github.com/MetricsDI/DIMetrics

Tools

LayoutLM is a BERT-like transformer model. We employ sequence labeling approach with single Softmax classifier after the encoder, and train over approximately 18,000 internal proprietary invoices using cross entropy loss function. To group nested line-item classifications, we use Probabilistic Soft Logic (PSL) **[1]** to classify parent line item IDs.
DeepCPCFG uses an expert-provided grammar and language model potentials as rules, operating on two-dimensional sequences formed by a directed graph representation of the page structure **[2]**. Unlike *LayoutLM*, *DeepCPCFG* does not require bounding box labels, but uses ground truth key-value pairs as inputs, and latently learns the mapping to bounding boxes.
Microsoft Form Recognizer is used as an industry benchmark end-to-end model, accessible via API calls. We benchmark the pre-built receipt model. We do share results of training a custom model on *CORD* data, due to inability to create custom parent-child predictions with the API.

Conclusion

We have shared *DI-Metrics*[a], a library for objective evaluation of IE Document Intelligence Tasks. The library provides a comprehensive set of metrics for use by researchers and industry practitioners to use and transparently benchmark information extraction models. We also introduced **UHED** metric.

References

[1] Nigel P. Duffy and Sai Akhil Puranam and Sridhar Dasaratha and Karmvir Singh Phogat and Sunil Reddy Tiyyagura *DeepPSL: End-to-end perception and reasoning with applications to zero shot learning*, arXiv, 2021

[2] Chua, Freddy C. and Duffy, Nigel P. *DeepCPCFG: Deep Learning and Context Free Grammars for End-to-End Information Extraction*, Document Analysis and Recognition – ICDAR 2021

[3] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li and Ming Zhou *DocBank: A Benchmark Dataset for Document Layout Analysis*, CoRR Journal, 2020

[4] Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang and Minjoon Seo *Spatial Dependency Parsing for 2D Document Understanding*, CoRR Journal, 2020

[5] Jonker, Roy and Volgenant, Ton *Improving the Hungarian assignment algorithm*, Operations Research Letters, 1986

[6] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang and Lidong Zhou *LayoutLMv2: Multimodal Pre-training for Visually-Rich Document Understanding* arXiv 2020

Disclaimer

The views reflected in this article are the views of the authors and do not necessarily reflect the views of the global EY organization or its member firms.

Metrics

We provide a library to ease consistent comparison of VRD model performance on IE tasks. The library is a collection of existing and new IE metrics (Table below) accessible through a Python3 API. Many metrics are dynamic programs based on edit distance, and they are known to be computationally expensive. Our implementations are accelerated by pre-compilation in Cython. We also introduce a novel metric for handling evaluation of hierarchical fields, **Unordered Hierarchical Edit Distance (UHED)**.

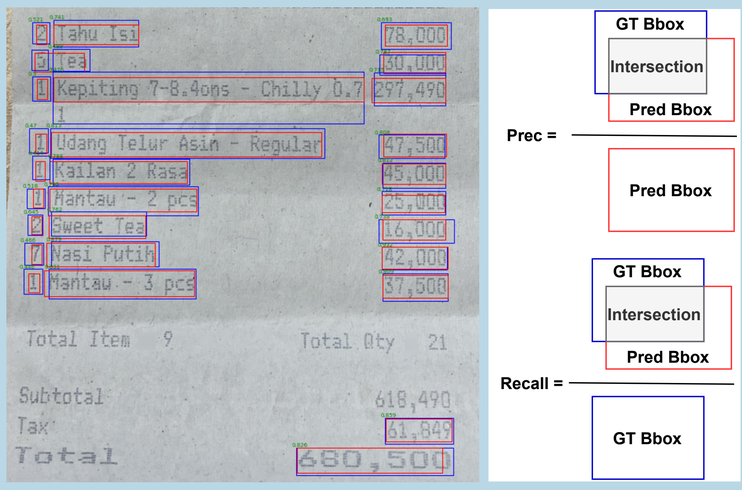
Metric Type	Metrics Name	Range
Text-Based (Field Level)	Exact Match	*True, False*
	Raw Levenshtein Distance	*0 - min(GT, P)*
	Raw Longest Common Subsequence (LCSeq)	*0 - min(GT, P)*
	Token Classification	*0 - 1*
Geometric-Based (Field Level)	Grouped Bbox by class IoU (IoU_G)	*0 - 1*
	Constituent Bbox by class IoU (IoU_C)	*0 - 1*
Hierarchical (Document Level)	Hierarchical Edit Distance (HED)	*0 - 1*
	Unordered Hierarchical Edit Distance (UHED)	*0 - 1*

Text-Based metrics

Text-based metrics measure the presence of typing or spelling errors and access the convergence of two strings. In **Exact Match (EM)** metric, we simply check whether the entire predicted string *P* is exactly the same as the ground truth string *GT*. **Levenshtein Edit Distance (LED)** between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. The **Longest Common Subsequence (LCSeq)** is the minimum number of insertions and deletions required to change one string to the other.

Geometric-Based metrics

Grouped Bbox by class (IoU_G) approach one computes the overlap of aggregated boxes by calculating a convex-hull minimal spanning box of all constituent bounding boxes surrounding the entire field and thus include any spaces between constituent OCR as well. **Constituent Bbox by class (IoU_C)** is adapted from the DocBank dataset paper **[3]**, where instead of taking the area of the entire field, we only consider the areas of individual tokens (words).



Left (invoice) - receipt ground truths (red), predictions (blue)
Right - geometrical explanation of IoU (precission and recall)

Hierarchical metrics

Hierarchical Metrics are applied when the fields of interest are nested. In **[4]**, edit distances are extended from strings to table cells of strings, using a tree-based edit distance for table cell recognition. **Hierarchical Edit Distance (HED)** was proposed in **[1]**. This metric also covers information about non-nested and hierarchical fields (line-items), effectively only requiring that the ordering of line-items within a document and words within a field remain the same, while the ordering of fields within a line-item may be permuted without impacting the distance. Our proposed **Unordered Hierarchical Edit Distance (UHED)** relaxes HED, allowing unordered lists of line-items. We apply Hungarian **[5]** assignment algorithm to find the optimal (*GT*, *P*) pairs by minimizing the matrix of input distances for each possible candidate p

Experimental Results

To test application of the metrics on models and data, we use CORD Receipts dataset. In Table below we present a comparison of HED and UHED metrics for three models: LayoutLM Base V1 **[6]**, DeepCPCFG and Microsoft Form Recognizer pre-built receipt model.

	HED			UHED		
CORD	**F1**†	**Precision**†	**Recall**†	**F1**†	**Precision**†	**Recall**†
LayoutLM + PSL LI Rules	0.89	0.88	0.91	0.92	0.92	0.94
LayoutLM + Simple LI Rule	0.86	0.85	0.89	0.92	0.96	0.90
DeepCPCFG	**0.96**	**0.97**	**0.97**	**0.97**	**0.98**	**0.97**
MSFT Form Recognizer	0.81	0.91	0.75	0.85	0.96	0.78