DocILE 2023 Teaser: Document Information Localization and Extraction

Štěpán Šimsa[1(✉)] , Milan Šulc[1] , Matyáš Skalický[1] , Yash Patel[2] ,
and Ahmed Hamdi[3]

[1] Rossum.ai, Prague, Czech Republic
{stepan.simsa,milan.sulc,matyas.skalicky}@rossum.ai
[2] Visual Recognition Group, Czech Technical University in Prague,
Prague, Czech Republic
patelyas@fel.cvut.cz
[3] University of La Rochelle, La Rochelle, France
ahmed.hamdi@univ-lr.fr
http://www.rossum.ai/

Abstract. The lack of data for information extraction (IE) from semi-structured business documents is a real problem for the IE community. Publications relying on large-scale datasets use only proprietary, unpublished data due to the sensitive nature of such documents. Publicly available datasets are mostly small and domain-specific. The absence of a large-scale public dataset or benchmark hinders the reproducibility and cross-evaluation of published methods. The DocILE 2023 competition, hosted as a lab at the CLEF 2023 conference and as an ICDAR 2023 competition, will run the first major benchmark for the tasks of *Key Information Localization and Extraction* (KILE) and *Line Item Recognition* (LIR) from business documents. With thousands of annotated real documents from open sources, a hundred thousand of generated synthetic documents, and nearly a million unlabeled documents, the DocILE lab comes with the largest publicly available dataset for KILE and LIR. We are looking forward to contributions from the Computer Vision, Natural Language Processing, Information Retrieval, and other communities. The data, baselines, code and up-to-date information about the lab and competition are available at https://docile.rossum.ai/.

Keywords: Information extraction · Dataset · Benchmark · KILE · LIR · Business documents · Natural language processing · Computer vision

1   Introduction

The majority of business-to-business (B2B) communication takes place through the exchange of semi-structured documents such as invoices, purchase orders, and delivery notes. Information from the documents is typically extracted by humans and entered into information systems. This process is time-consuming, expensive, and repetitive. Automating the information extraction process has the potential to considerably reduce manual human labor, allowing people to focus on more creative and strategic tasks.

Fig. 1. Example invoice with annotations of fields and line items (LIs). Categories of fields within the LI are depicted by their color. Values of fields in the LI are not visualized in the Figure, but they are annotated in the dataset.

Automating document information extraction is challenging because semantic and syntactic understanding is required. These documents are designed to be interpretable by humans, not machines. An example with semantic information is shown in Fig. 1. Information extraction approaches must handle varying layouts, semantic fields and multiple input modalities at the intersection of computer vision, natural language processing and information retrieval. While there has been progress on the task [4,7,14,15,18,19,25,34], there is no publicly available large-scale benchmark to train and compare these approaches, an issue that has been noted by several authors [5,16,24,26,29]. Existing approaches are trained on privately collected datasets, hindering their reproducibility, fair comparisons and tracking field progression [11,23,24].

To mitigate the aforementioned issues, the DocILE lab provides a public research benchmark on a large-scale dataset. This benchmark was built by knowing the domain- and task-specific aspects of business document information localization and extraction. The DocILE benchmark will allow cross-evaluation and enable the reproducibility of experiments in business document information extraction. The dataset is the largest public source of densely annotated business documents. It consists of 8715 annotated pages of 6680 real business documents along with 100,000 synthetic documents and 3.4 million unlabeled pages of nearly a million real business documents. To mimic the real world use case, the dataset emphasizes layout diversity and contains over a thousand unique layouts. With the large amount of diverse documents and high-quality annotations, the dataset will allow researchers to investigate different aspects of document information extraction, including supervised, semi-supervised and unsupervised learning and domain adaptation.

2   Dataset and Tasks

The DocILE benchmark comes with a *labeled* dataset of 6680 documents from publicly available sources which were manually annotated for the tasks of *Key Information Localization and Extraction* and *Line Item Recognition*, described below in Sects. 2.2 and 2.3 respectively. Additionally, we provide a set of 100K *synthetic* documents generated with the task annotations and 932K *unlabeled* documents, as both synthetic training data [6,10,22] and unsupervised pre-training [33] have demonstrated to aid machine learning in different domains.

2.1   Dataset Characteristics

Table 1 shows the size of the challenge dataset. All documents in the dataset were classified[1] as invoice-like documents (i.e., tax invoice, order, proforma invoice) by a model pre-trained on a private dataset. Additionally, in the labelling process, documents misclassified as invoice-like were manually removed from the dataset (e.g., budgets or financial reports, as such document types contain different information than standard invoice-like documents).

To ensure a high variance of document layouts in the dataset, *unlabeled* documents were clustered into layouts[2]. Only a limited number of documents per layout were selected for annotation. The clustering is based on the location of field detections[3] predicted by a proprietary model for KILE pre-trained on a private dataset. Furthermore, to encourage solutions that generalize well to previously unseen layouts, the train./val./test split is done such that the validation and test sets contain layouts unseen in the training set (to measure the model's generalization) as well as some seen layouts (in practice, it is common to observe known layouts and important to read them out perfectly). Meta-information describing the layouts is included in the dataset annotations. The *synthetic* documents were generated using an unpublished rule-based document synthesizer based on layout annotations of 100 documents from the *labeled* set.

The dataset will be shared in the form of pre-processed[4] document PDFs with task annotations in JSON. As an additional resource, we will also provide predictions of text tokens (using OCR) including the location and text of the detected tokens.

---

[1] Using a proprietary document type classifier from Rossum.ai.

[2] We loosely define layout as the positioning of fields of each type in a document. Rather than requiring absolute positions to match perfectly, we allow transformations caused by different length of values, translations of whole sections (e.g. vertical shift caused by different lengths of tables) and translation, rotation and scaling of the whole document.

[3] The distance used for clustering relates to the difference in the relative x-translations between pairs of fields within a document. Vertical shifts are not penalized, since they commonly appear among documents of the same layout.

[4] Pre-processing consists of correcting page orientation, fixing or discarding broken pdfs and of de-skewing scanned documents and normalizing them to 150 DPI.

Table 1. Overview of the three parts of the challenge dataset.

	Labeled	Synthetic	Unlabeled
Documents	6680	100 000	932 467
Pages	8715	100 000	3.4M
Layout clusters	1152	100	*Unknown*
Pages per doc.	1–3	1	1–884

The data was sourced from two public data sources: UCSF Industry Documents Library [30] and Public Inspection Files (PIF) [32]. The UCSF Industry Documents Library contains documents from industries that influence public health, such as tobacco companies. The majority of the documents are from the 20th century. This source was previously used to create document datasets: RVL-CDIP [12] (subset of IIT-CDIP [17] and superset of FUNSD [9]) and DocVQA [21]. Filters in the UCSF public API [31] were used to retrieve only publicly available invoice-like documents with at most 3 pages, no redacted information and a threshold on document date[5]. PIF contains documents (invoices, orders, "contracts") from TV and radio stations for political campaign ads. This source was previously used to create Deepform [28].

2.2   Track 1: Key Information Localization and Extraction

The goal of the first track is to localize key information of pre-defined categories (field types) in the document. It is derived from the task of *Key Information Localization and Extraction* (KILE), as defined in [26].

KILE extends the common definition of Key Information Extraction (KIE) by additionally requiring the location of the extracted information within the document. Such annotation is missing even in the KIE datasets [3,27]. While localization is typically not needed at the end of document processing, it plays a vital role in applications that require human validation, and it is a valuable form of supervision for vision-based methods. Compared to *Semantic Entity Recognition*, as defined by [33], bounding boxes in KILE are not limited to individual words (tokens).

We focus the challenge on detecting semantically important values corresponding to tens of different field types rather than fine-tuning the underlying text recognition. Towards this focus, we provide word-level text detections for each document, we choose an evaluation metric (below) that doesn't pay attention to the text recognition part, and we simplify the task in the challenge by only requiring correct localization of the values in the documents in the primary metric. Text extractions are checked besides the locations and field types in a separate evaluation (the leaderboard ranking does not depend on it) and any post-processing of values (deduplication, converting dates to a standardized

---

[5] Old documents from this source are not included, since e.g. typewriter documents differ from today's document distribution.
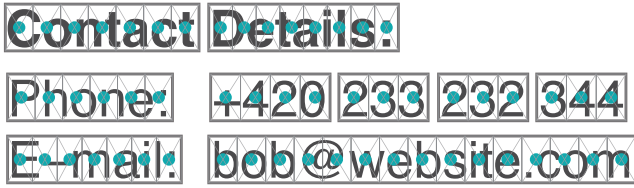
Fig. 2. Each word is split uniformly into pseudo-character boxes based on the number of characters. Pseudo-Character Centers are the centers of these boxes.
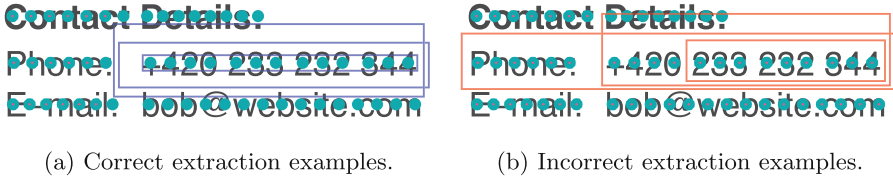

(a) Correct extraction examples.   (b) Incorrect extraction examples.

Fig. 3. Visualization of correct and incorrect bounding box predictions to capture the email address. Bounding box must include exactly the Pseudo-Character Centers that lie within the ground truth annotation. Note: In 3a, only one of the predictions would be considered correct if all three boxes were predicted.

format etc.) that is otherwise needed in practice is omitted. With the simplifications, the main task can also be viewed as a detection problem.

Challenge Evaluation Metric: Since the task is framed as a detection problem, the standard *Average Precision* metric will be used as the main evaluation metric. Unlike the common practice in object detection, where true positives are determined by thresholding the Intersection-over-Union, we use a different criterion tailored to better evaluate the usefulness of detections for text read-out. Inspired by the CLEval metric [1] used in text detection, we measure whether the predicted area contains all related character centers (and none others). Since the character-level annotations are hard to obtain, we use CLEval's definition of Pseudo-Character Center (PCC) (see Fig. 2). See Fig. 3 for examples of correct and incorrect detections.

Beyond the challenge leaderboard based on the metric described above, we set up a secondary benchmark for end-to-end KILE, where a correctly recognized field also needs to exactly read out the text. We invite all participants to provide the text value predictions, but it is not required for challenge submissions.

2.3   Track 2: Line Item Recognition

The goal of the second track is to localize key information of pre-defined categories (field types) and group it into line items [2,4,13,20,24]. A *Line Item* (LI) is a tuple of fields (i.e., *description*, *quantity*, and *price*) describing a single object instance to be extracted, e.g., a row in a table, as visualized in Fig. 1.

This track is derived from the task of *Line Item Recognition* (LIR) [26] and is related to *Table Understanding* [13] and *Table Extraction* [8, 35]—problems where the tabular structure is also crucial for IE. Unlike these tasks, LIR does not explicitly rely on the structure but rather reflects the information to be extracted and stored.

Challenge Evaluation Metric: The main evaluation metric is the micro F1 score over all line item fields. A predicted line item field is correct if it fulfills the requirements from Track 1 (on field type and location) and if it is assigned to the correct line item. Since the matching of ground truth (GT) and predicted line items may not be straightforward due to errors in prediction, our evaluation metric chooses the best matching in two steps:

1. For each pair of predicted and GT line items, the predicted fields are evaluated as in Track 1.
2. Find the maximum matching between predicted and GT line items, maximizing the overall recall.

Similarly to the previous track, an out-of-competition end-to-end benchmark will assess the correctness of the extracted text values.

3   Conclusions

The first edition of the DocILE lab at CLEF 2023 and the ICDAR 2023 Competition on Document Information Localization and Extraction will present the largest benchmark for information extraction from semi-structured business documents, and will consist of two tasks: *Key Information Localization and Extraction (KILE)* and *Line Item Recognition (LIR)*. Participants will be given a collection of thousands of labeled documents, together with a hundred thousand of synthetic documents and nearly a million unlabeled real documents that can be used for unsupervised pre-training.

This Teaser paper summarizes the motivation and the main characteristics of the tasks. Given the input documents are practically a combination of visual- and text- inputs, we are looking forward to the contributions of several communities, including *Information Retrieval*, *Natural Language Processing*, and *Computer Vision*.

To access the data, the repository, baseline implementations, and updates regarding the challenge, please refer to https://docile.rossum.ai/.

References

1. Baek, Y., et al.: CLEVAL: character-level evaluation for text detection and recognition tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 564–565 (2020)

2. Bensch, O., Popa, M., Spille, C.: Key information extraction from documents: evaluation and generator. In: Abbès, S.B., et al. (eds.) Proceedings of DeepOntoNLP and X-SENTIMENT. CEUR Workshop Proceedings, vol. 2918, pp. 47–53. CEUR-WS.org (2021)

3. Borchmann, Ł., et al.: DUE: end-to-end document understanding benchmark. In: Proceeedings of NeurIPS (2021)

4. Denk, T.I., Reisswig, C.: Bertgrid: contextualized embedding for 2D document representation and understanding. arXiv preprint arXiv:1909.04948 (2019)

5. Dhakal, P., Munikar, M., Dahal, B.: One-shot template matching for automatic document data capture. In: Proceeedings of Artificial Intelligence for Transforming Business and Society (AITB), vol. 1, pp. 1–6. IEEE (2019)

6. Dosovitskiy, A., et al.: Flownet: learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2758–2766 (2015)

7. Garncarek, Ł, et al.: LAMBERT: layout-aware language modeling for information extraction. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12821, pp. 532–547. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86549-8_34

8. Göbel, M.C., Hassan, T., Oro, E., Orsi, G.: ICDAR 2013 table competition. In: Proceedings of ICDAR, pp. 1449–1453. IEEE Computer Society (2013). https://doi.org/10.1109/ICDAR.2013.292

9. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: a dataset for form understanding in noisy scanned documents. In: Accepted to ICDAR-OST (2019)

10. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315–2324 (2016)

11. Hamdi, A., Carel, E., Joseph, A., Coustaty, M., Doucet, A.: Information extraction from invoices. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12822, pp. 699–714. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86331-9_45

12. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR) (2015)

13. Holeček, M., Hoskovec, A., Baudiš, P., Klinger, P.: Table understanding in structured documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 5, pp. 158–164. IEEE (2019)

14. Holt, X., Chisholm, A.: Extracting structured data from invoices. In: Proceedings of the Australasian Language Technology Association Workshop 2018, pp. 53–59 (2018)

15. Katti, A.R., et al.: Chargrid: towards understanding 2D documents. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018, pp. 4459–4469. Association for Computational Linguistics (2018). https://aclanthology.org/D18-1476/

16. Krieger, F., Drews, P., Funk, B., Wobbe, T.: Information extraction from invoices: a graph neural network approach for datasets with high layout variety. In: Ahlemann, F., Schütte, R., Stieglitz, S. (eds.) WI 2021. LNISO, vol. 47, pp. 5–20. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86797-3_1

17. Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., Heard, J.: Building a test collection for complex document information processing. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 665–666 (2006)

18. Lin, W., Gao, Q., Sun, L., Zhong, Z., Hu, K., Ren, Q., Huo, Q.: ViBERTgrid: a jointly trained multi-modal 2D document representation for key information extraction from documents. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12821, pp. 548–563. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86549-8_35

19. Liu, W., Zhang, Y., Wan, B.: Unstructured document recognition on business invoice. Mach. Learn., Stanford iTunes Univ., Stanford, CA, USA, Technical Report (2016)

20. Majumder, B.P., Potti, N., Tata, S., Wendt, J.B., Zhao, Q., Najork, M.: Representation learning for information extraction from form-like documents. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, pp. 6495–6504 (2020). https://doi.org/10.18653/v1/2020.acl-main.580

21. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: a dataset for VQA on document images. In: Proceedings of WACV, pp. 2199–2208. IEEE (2021). https://doi.org/10.1109/WACV48630.2021.00225

22. Nassar, A., Livathinos, N., Lysak, M., Staar, P.W.J.: Tableformer: table structure understanding with transformers. CoRR abs/2203.01017 (2022). https://doi.org/10.48550/arXiv.2203.01017

23. Palm, R.B., Laws, F., Winther, O.: Attend, copy, parse end-to-end information extraction from documents. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 329–336. IEEE (2019)

24. Palm, R.B., Winther, O., Laws, F.: Cloudscan - a configuration-free invoice analysis system using recurrent neural networks. In: Proceedings of ICDAR, pp. 406–413. IEEE (2017). https://doi.org/10.1109/ICDAR.2017.74

25. Schuster, D., et al.: Intellix-end-user trained information extraction for document archiving. In: 2013 12th International Conference on Document Analysis and Recognition, pp. 101–105. IEEE (2013)

26. Skalický, M., Šimsa, Š., Uřičář, M., Šulc, M.: Business document information extraction: towards practical benchmarks (2022). https://doi.org/10.48550/ARXIV.2206.11229, https://arxiv.org/abs/2206.11229

27. Stanisławek, T., et al.: Kleister: key information extraction datasets involving long documents with complex layouts. In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR 2021. LNCS, vol. 12821, pp. 564–579. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86549-8_36

28. Stray, J., Svetlichnaya, S.: Deepform: Extract information from documents (2020). https://wandb.ai/deepform/political-ad-extraction, benchmark

29. Sunder, V., Srinivasan, A., Vig, L., Shroff, G., Rahul, R.: One-shot information extraction from document images using neuro-deductive program synthesis. arXiv preprint arXiv:1906.02427 (2019)

30. Web: Industry Documents Library. https://www.industrydocuments.ucsf.edu/. Accessed 20 Oct 2022

31. Web: Industry Documents Library API. https://www.industrydocuments.ucsf.edu/research-tools/api/, Accessed 20 Oct 2022
32. Web: Public Inspection Files. https://publicfiles.fcc.gov/. Accessed 20 Oct 2022
33. Xu, Y., et al.: LayoutXLM: multimodal pre-training for multilingual visually-rich document understanding. CoRR (2021)
34. Zhao, X., Wu, Z., Wang, X.: CUTIE: learning to understand documents with convolutional universal text information extractor. CoRR abs/1903.12363 (2019). https://arxiv.org/abs/1903.12363
35. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global table extractor (GTE): a framework for joint table identification and cell structure recognition using visual context. In: Proceedings of WACV, pp. 697–706. IEEE (2021). https://doi.org/10.1109/WACV48630.2021.00074