DoSA : A System to Accelerate Annotations on Business Documents with Human-in-the-Loop

Neelesh K Shukla and **MSP Raja** and **Raghu Katikeri** and **Amit Vaid**
State Street Corporation
Bengaluru, KA, India
{nshukla, smaddila1, rkatikeri, avaid}@statestreet.com

Abstract

Business documents come in a variety of structures, formats and information needs which makes information extraction a challenging task. Due to these variations, having a document generic model which can work well across all types of documents and for all the use cases seems far-fetched. For document-specific models, we would need customized document-specific labels. We introduce DoSA (**Do**cument **S**pecific **A**utomated Annotations), which helps annotators in generating initial annotations automatically using our novel bootstrap approach by leveraging document generic datasets and models. These initial annotations can further be reviewed by a human for correctness. An initial document-specific model can be trained and its inference can be used as feedback for generating more automated annotations. These automated annotations can be reviewed by human-in-the-loop for the correctness and a new improved model can be trained using the current model as pre-trained model before going for the next iteration. In this paper, our scope is limited to Form like documents due to limited availability of generic annotated datasets, but this idea can be extended to a variety of other documents as more datasets are built. An open-source ready-to-use implementation is made available on GitHub. [1]

1 Introduction

With the recent advancements in technology and increased adoption of digitization, almost all organizations maintain and exchange business documents in digitized formats like PDFs, scans, faxes, images etc. These documents come in all shape, sizes and format like invoices, emails, medical reports, contracts, scientific papers and many more. The majority of the research has concentrated on documents present on the web that do not adequately capture the complexity of analysis or comprehension required for business documents. These documents require a multidisciplinary approach that includes understanding of layout and structure, computer vision, natural language processing. Usually, organizations rely on humans to manually process these documents. The ability to read, understand and interpret these documents is referred as Document Intelligence (DI) or Document Understanding. There have been recent advancements in this area specifically with deep learning where many architectures (Huang et al., 2022; Appalaraju et al., 2021; Kim et al., 2021) and annotated datasets (Mathew et al., 2020; Zhong et al., 2019; Huang et al., 2019; Park et al., 2019) have been published for various DI tasks like Document Classification, Document Visual Q&A, Form Understanding etc. These publicly annotated datasets mostly capture only few document types like receipts, scientific articles etc. It becomes challenging & tedious to have annotated public datasets available for various document types which therefore brings in the very need of having customized annotated datasets for specific use cases and document types.

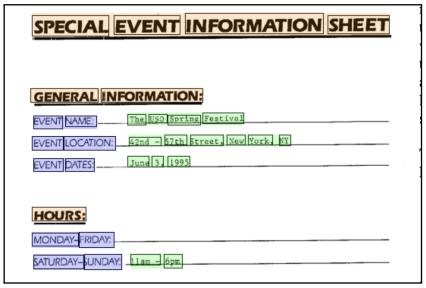

Figure 1: Form Example from FUNSD: Keys are represented in blue, headers in orange, and Values in green.

To limit our scope, in this paper we are focusing on information extraction from documents that follow a form-like structure. Forms are documents that have information usually present in Question-Answer or Key-Value format as shown in Figure 1. Documents like invoices, driving licenses, passports, medical records, financial statements, tax

[1] https://github.com/neeleshkshukla/DoSA

Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances), pages 23 - 27

forms, quotations, payment cards, etc. fall under this category.

There has been an effort in building a generic form dataset FUNSD[2] (Jaume et al., 2019). FUNSD is a dataset for form understanding in noisy scanned documents that aim at extracting and structuring the textual content of forms. It proposes an idea where a generic document can be represented via generic information and labels like question (or key), answer (or value), header and others. A model can consume this kind of generic representation to extract generic information. In most of the scenarios, users are interested in document-specific meaningful labels like document number, document date, etc and extracting a subset of information. Having a document generic labeling approach results in a noisy and verbose extraction. Therefore a need for document-specific annotations arises. Commercial solutions like Microsoft Form Recognizer[3] and Google Document AI [4] mostly support specific document type prebuilt models and provide a facility to custom train a model for specific documents[5] which require new annotations. The SOTA models of document intelligence use multimodality of the document: text, position and image. Due to a change in the format or layout of the document, these modalities might be affected and a new model needs to be trained which will need a new set of annotations. With these, manual annotation becomes repetitive, laborious, expensive and time-consuming.

To reduce the time and human effort, we are proposing an active learning based automated annotation system DoSA (**Do**cument **S**pecific **A**utomated Annotations), where the initial set of document-specific annotations are generated by the system which can be reviewed by human annotators for correctness. An initial model can be trained with these annotations and its inference can be taken by the system as feedback to generate annotations on new documents and improve the model incrementally with the human in the loop.

The main contribution of this paper is a novel bootstrapping approach to generate automated document-specific labels. To the best of our knowledge, this is the first attempt at generating automated annotations on business documents that contains visual and layout structure information along with the text. All other previous approaches have mostly focused on web or text (Dill et al., 2003; Wilson et al., 2018) or images (Yu et al., 2019). We have seen an attempt on research documents but the scope was limited to automatically annotating documents with topic (Singhal et al., 2013).

2 DoSA System

A high-level flow of the DoSA system is shown in Figure 2, where initial document-specific annotations are generated by document-generic model which is later reviewed by a human for correctness. With these initial annotations, a document-specific model is trained and its inference is taken as feedback to annotate further documents. A human will again review and correct the annotations and the reviewed documents can be added back to training for further improving the model. As the model matures, eventually the user would end up correcting a minimal number of annotated fields/documents. The journey for the model to get more precise with less human feedback is achieved by employing active learning strategies like uncertainty-based sampling which improve the document-specific model performance in very few iterations.

2.1 Bootstrapping: Generating Initial Document Specific Annotations with Document Generic Model

Here are the steps that are followed in generating annotations on the initial set of documents:

- A document is processed via an OCR engine to get the words and their respective bounding boxes. DoSA uses open source OCR engine pytesseract[6].
- As an intermediate step, the text areas are identified as 'Keys' and 'Values' in these documents (Section 2.1.1).
- Link respective Key and Values and form a key-value pair <K, V> (Section 2.1.2).
- Document-specific annotations can be generated by labeling the area identified as value V with the text of the area identified by the respective key K (Section 2.1.3).

[2]https://guillaumejaume.github.io/FUNSD/

[3]https://azure.microsoft.com/en-in/services/form-recognizer/

[4]https://cloud.google.com/document-ai

[5]https://docs.microsoft.com/en-us/azure/applied-ai-services/form-recognizer/concept-custom

[6]https://pypi.org/project/pytesseract/

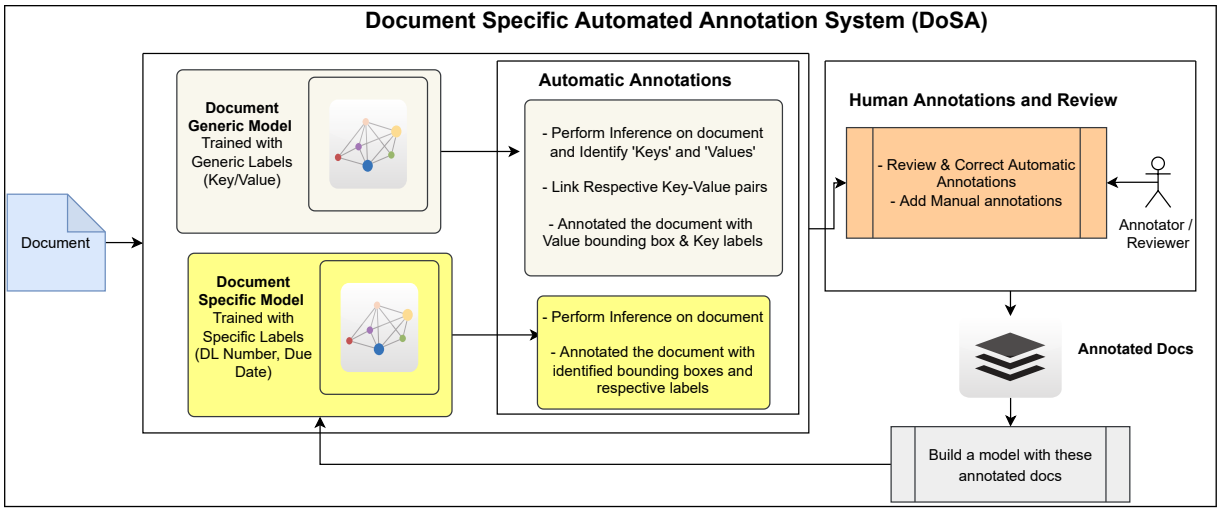

Figure 2: DoSA System Overview: Generating Annotations with Human in the Loop

2.1.1 Locating Keys and Values

As an intermediate step, text areas in a document are identified as 'Key' and 'Value'. DoSA uses LayoutLMv3(Huang et al., 2022) model for entity/token classification fine-tuned on generic FUNSD dataset[7]. This can classify the word/token/entity in 'Key (Question)', 'Value (Answer)', 'Header' and 'Others' with F1 0.9078. The areas are represented by bounding box coordinates <x1,y1, x2, y2> which are used for comparing the position and drawing rectangles in the next sections. In the fax cover example shown in figure 3, 'To', 'Fax Number', 'Phone Number', 'Date' has been identified as keys and 'George Baroody', '(336) 335-7392', '12/10/98' as values.

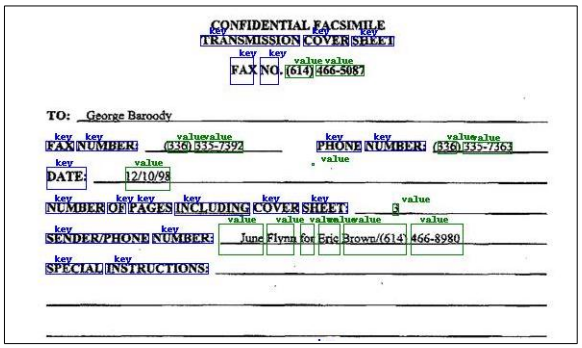

Figure 3: Areas marked with generic 'Key' and 'Value' labels as described in section 2.1.1

2.1.2 Key and Value Pair Linking

After identifying keys and values, DoSA links respective pairs <K, V>. There have been multiple recent works addressing this Form Entity Linking problem (Li et al., 2021; Zhang et al., 2021). These works have F1 0.4 and 0.64 respectively. These SOTA models are not good enough and were resulting in a lot of noisy pairs. Based on manual observations of a few documents, we designed the following heuristics to identify key-value pairs.

Given a list of values V ordered by their position in a document, Value Vj is linked to candidate key Ki if it satisfies the following conditions:

H1: Position of Ki is less than the position of Vj.

H2: Ki is not linked to any other Value Vk

H3: Ki is the closest to Vj compared to other candidate keys Km which satisfy H1 and H2.

If No such key is found for a Value Vj. Vj will be dropped else pair <Ki, Vj> is added to the output. For the document shown in figure 3, some of the examples are <Fax Number: (336) 335-7392> and <Date: 12/10/98>.

2.1.3 Document Specific Annotations and Review with Human-in-the-Loop

Once the <Key, value> pairs are identified, annotations can be generated by drawing the bounding boxes around 'value' and annotating it with the text of the respective 'key'. These automated annotations can be submitted for review and modifications with Human-in-the-loop.

An example is shown in Figure 4, once the key-value pair <Fax Number: (336) 335-7392> identified, the value (336) 335-7392 has been annotated with respective key 'Fax Number'.

[7]https://huggingface.co/nielsr/layoutlmv3-finetuned-funsd

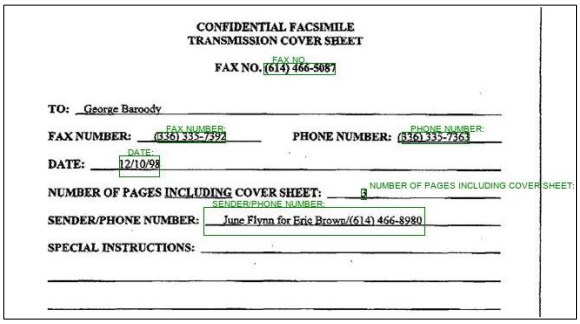

Figure 4: Document specific annotations by labeling regions/texts identified as 'Value' with respective 'Key' in intermediate annotations as shown in figure 3.

2.2 Annotations with Document Specific Model

A custom initial model can be built by fine-tuning the generic model used in section 2.1.1 on these reviewed initial annotations which now have document-specific labels. This document-specific model can be used to generate annotations via inference feedback on new documents. As more documents and annotations are added and reviewed, the model will eventually get mature.

3 Conclusion and Future Work

In this work, we presented DoSA, a system to generate document specific annotations from model built on document generic datasets. Our scope was limited to Form like document which can be further enhanced with the availability of new type of generic datasets. This system in current state can only take one type of document to generate one set of annotations. In case the users have multiple type of documents, they have to group the documents by type beforehand and use this system for individual groups. A layer can be added on top of DoSA system to automatically classify the documents and use DoSA for individual groups. As this work is still in progress, in this paper we focused on proposing this idea. We are planning to discuss the effectiveness of our proposed approaches and overall system in the near future.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 973–983.

Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. 2003. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. WWW '03, page 178–186, New York, NY, USA. Association for Computing Machinery.

Yupang Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACM Multimedia 2022*.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. Donut: Document understanding transformer without ocr. *CoRR*, abs/2111.15664.

Yulin Li, Yuxi Qian, Yuchen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021. Structext: Structured text understanding with multi-modal transformers. *CoRR*, abs/2108.02923.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for VQA on document images. *CoRR*, abs/2007.00398.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. {CORD}: A consolidated receipt dataset for post-{ocr} parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Ayush Singhal, Ravindra Kasturi, and Jaideep Srivastava. 2013. Automating document annotation using open source knowledge. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 199–204.

Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, and Noah A. Smith. 2018. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Trans. Web*, 13(1).

Chao-Wei Yu, Yen-Lin Chen, Ko-Feng Lee, Chen-Hsiang Chen, and Chia-Yu Hsiao. 2019. Efficient intelligent automatic image annotation method based

on machine learning techniques. In *2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, pages 1–2.

Yue Zhang, Zhang Bo, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. 2021. Entity relation extraction as dependency parsing in visually rich documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2759–2768, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.