MexPub: Deep Transfer Learning for Metadata Extraction from German Publications

Zeyd Boukhers Nada Beili Timo Hartmann Prantik Goswami Muhammad Arslan Zafar
Institute for Web Science and Technologies (WeST)
University of Koblenz-Landau
Koblenz, Germany
{boukhers,nbeili,tihartmann,prantik,arslanzafar}@uni-koblenz.de

***Abstract*—In contrast to most of the English scientific publications that follow standard and simple layouts, the order, content, position and size of metadata in German publications vary greatly among publications. This variety makes traditional NLP methods fail to accurately extract metadata from these publications. In this paper, we present a method that extracts metadata from PDF documents with different layouts and styles by viewing the document as an image. We used Mask R-CNN which is trained on COCO dataset and finetuned with PubLayNet dataset that consists of 200K PDF snapshots with five basic classes (e.g. text, figure, etc). We refine-tuned the model on our proposed synthetic dataset consisting of 30K article snapshots to extract nine patterns (i.e. author, title, etc). Our synthetic dataset is generated using contents in both languages German and English and a finite set of challenging templates obtained from German publications. Our method achieved an average accuracy of around 90% which validates its capability to accurately extract metadata from a variety of PDF documents with challenging templates.**

***Index Terms*—transfer learning, metadata extraction, neural networks**

I. Introduction

The availability and accessibility of academic metadata allow the development of semantic-enable services, such as authors' profiling, bibliometrics, and scientific social network analysis. However, still, a significant part of bibliographic data in disciplines such as social science is not accessible via bibliographic databases and a vast amount of already existing scientific documents have incomplete or entirely missing metadata information [13].

One way to overcome this problem is by a post hoc processing which is automatically extracting metadata from PDF documents. Consequently, several approaches for automatic metadata extraction from scientific documents [6], [8], [13]. have been proposed. These approaches are used by different digital libraries and publishers due to their stable accuracy when the layout and the structure of the PDF documents are standard. However, in a comparative evaluation of several metadata extraction tools including GROBID [8], Lipinski et al. [6] found that while some tools achieve an accuracy of around 90% on the title, they can extract abstract or date with accuracy only between 26% and 75%. Moreover, state-of-the-art methods do not provide high accuracy when applied to non-English layouts, such as those of German scientific publications. This is confirmed by the absence of a lot of German publications from bibliographic indices [3]. Recently several approaches have been proposed to process publications in German language [1].

Over the past years, deep learning methods have dramatically improved the state-of-the-art in many domains, such as object recognition and object detection. In line with Stahl et al. [11], we assume in this paper that scientific publications have an inherent structure that can be learnt by a deep learning network to distinguish among metadata patterns. Hence, this paper proposes a novel deep learning method for extracting metadata from scientific documents by treating PDF documents as images. The proposed method extracts information (e.g. author) by detecting its ROI in the image. To this end, we manually annotated a set of snapshots of each first page of 100 German scientific documents. Furthermore, we automatically generated a corpus consisting of around 44K annotated images of synthetic scientific papers based on the templates derived from the manual annotation. Using this dataset, we have fine-tuned an implementation of Mask R-CNN to distinguish between nine classes of metadata, namely; title, authors, journal, abstract, date, DOI, address, affiliation, and email addresses. The carried out evaluation demonstrates that our proposed model achieves a high average precision overall and across all classes. The main contributions of this paper are: I) Up to our knowledge, this is the first proposed model that tackles metadata extraction from PDF documents using a computer vision-based approach; II) Unlike conventional approaches, our model focuses on challenging and nonstandard layouts; III) We introduced a new and challenging synthetic dataset for metadata extraction; and IV) The experimental results on the proposed dataset demonstrate the effectiveness of *MexPub*.

II. Related Work

Throughout the past decades, several research works have investigated the problem of extracting metadata from scientific literature [13], [15], where most of them tackle this problem on a text level. Tkaczyk [13] suggests that information extraction from scientific documents usually consists of multiple subtasks: 1) parsing the input documents, 2) segmenting the content of a document into basic segments and determining their order within the document, and 3) detecting the attributes

2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL) | 978-1-6654-1770-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/JCDL52503.2021.00076

978-1-6654-1770-9/21/$31.00 ©2021 IEEE
DOI 10.1109/JCDL52503.2021.00076

of particular segments. The main problem of this multi-phase approach is aggregating errors over phases.

One standard way to easily and accurately extract information from documents is by relying on text and layout rules like in [10]. These approaches follow a set of predefined rules, which means that the layout of the paper has to be standard and known beforehand. Therefore, most of the earlier works addressed the problem of classifying segment strings in scientific documents using context-based classifiers such as Hidden Markov Models (HMMs) [12] and Conditional Random Fields (CRF) [8] or models like Support Vector Machines (SVMs) [2]. HMM models are known for their limitations when dealing with multiple non-independent features. SVMs, on the other hand, can handle a large variety of independent features [13], but lack the possibility of mapping a whole sequence of instances to a sequence of states.

As stated by Stahl et al. [11], more contemporary methods mostly utilize Conditional Random Fields (CRFs), which represent undirected graphical models trained to make inter-dependent predictions. In this way, these models combine the advantages of both finite-state HMM and discriminative SVM techniques by allowing arbitrary, dependent features and joint inference over entire sequences. Lipinski et al. [6] found that GROBID [8], a system using CRFs, performed significantly better than SVM-based approaches.

Despite the validated effectiveness of Neural Networks (NNs) in various machine learning tasks, in the field of metadata extraction, to date, only a limited number of approaches include deep learning techniques [7], [9], [11]. Siegel et al. [9] used distantly supervised NNs to extract figures from scientific documents, while Stahl et al. [11] propose a computer vision approach to recognise metadata patterns in images of scientific papers by modifying a version of Convolutional Neural Network (CNN) to classify content in scientific documents as one of two types, namely "paragraphs" and "non-paragraphs". Similarly, Liu et al. [7] implement a two-step deep learning approach using the image information of headers in scientific papers combined with the text content to classify the header's components into one of eight types of metadata.

III. Metadata Extraction from Publications (*MexPub*)

Our model is an implementation of Mask R-CNN, as proposed by He et al. [4]. Mask R-CNN extends Faster RCNN, which has two outputs for each candidate object, a class label, and bounding box offset. This extension is made by adding a branch for predicting an object mask and using Region of Interest (RoI)-Align instead of RoI-Pooling. The binary object mask represents the position on a pixel-level of each object within its bounding box.

We built our model using Detectron2 [1], which includes an implementation of MASK R-CNN models trained on the COCO dataset[2]. Specifically, we implemented Mask R-CNN with a ResNeXt [14] back-bone architecture with Feature Pyramid Network (FPN) following [5]. As Figure 1 illustrates, the model has three main blocks: (i) an FPN,(ii) a Regional Proposal Network, and (iii) RoI Heads. The first block consists of a *stem block* and four stages that contain multiple *bottleneck blocks*. More details can be found in the full version [3]

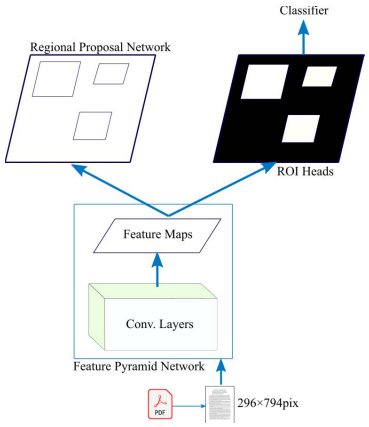

Fig. 1: Architecture of MASK R-CNN-FPN

For our task, we adopt the training configurations of the source model[4], which was originally fine-tuned on a total of 191,832 images stemming from the PubLayNet dataset [16]. PubLayNet consists of images of articles from PubMed Central™ Open Access (PMCOA) and each image is annotated with regions of the following five classes: title, text, list, table, and figure. This model is considered to be suitable for extracting metadata from scientific papers since (i) its backbone was trained on the very large COCO dataset, (ii) it was fine-tuned on a large data set of scientific document snapshots, which (iii) makes the task very similar to ours.

To adapt this model to the task of extracting metadata patterns from scientific documents, we, first, modified the last layer of the source model to output the nine target classes (i.e. title, authors, journal, abstract, date, DOI, address, affiliation, and email addresses) instead of the original five classes. The empirical experiments on a subset of 10^3 random samples from our training dataset demonstrates that the architecture with best performance is the one with 2 frozen layers and 15k iterations. Based on this finding, we refine-tuned the model using the full training dataset by setting the learning rate to 2.5×10^{-3}. The experiments were ran on Tesla P100 GPU using Google Colaboratory Pro[5].

IV. Experiments

We used a collection of 100 German scientific papers randomly selected from available publications in the SSOAR [6] repository. The dataset includes various scientific layouts, which allowed us to train *MexPub* on a diverse data set and

[1]https://github.com/facebookresearch/detectron2
[2]http://cocodataset.org/#home
[3]https://github.com/ZResearch/MexPub/blob/main/MexPub.pdf
[4]https://github.com/hpanwar08/detectron2
[5]https://colab.research.google.com
[6]https://www.gesis.org/ssoar/home

prevent the network from overfitting. In all the selected papers, most of the metadata is present on the first page. Therefore, we converted the first page of each paper (PDF) to JPEG. To speed up the training phase of *MexPub*, we resized all images to 596x794 pixels.

We manually annotated the regions corresponding to the desired metadata on a pixel-by-pixel level for each resized image. For this, we extended the metadata patterns proposed by Liu et al. [7]: title, author, affiliation, address, email, date, abstract, journal name, and DOI. We utilized the image annotation tool *Labelme*[7] to draw polygons on the metadata sections and saved the annotations in a COCO[8] format.

Due to the difficulty to annotate metadata on a pixel-by-pixel level, we extended our dataset by automatically generating synthetic papers based on the 28 most common layouts we identified during the mannual annotation phase. To this end, we randomly extracted metadata records from *SSOAR*[6], *DBLP* [9], and a list of scientific affiliations from Wikipedia[10]. For each of these layouts, we generated an average of 1600 synthetic papers by randomly inserting metadata from the extracted metadata at their corresponding positions on the first page. This process ensures that samples are not just duplicated with different content but that relatively different templates are generated. For example, when longer titles are broken over two or several lines, the template slightly changes from the original one. This change increases under other factors such as multiple authors instead of one, different address style or when the date is absent. The same same reprocessing procedure has been applied on the synthetic samples. The automatically generation of synthetic papers has expanded the dataset's size to around 44K papers with German and English content. More details about the annotation can be found in the full version 3^0.

To evaluate the model, we randomly split our training data into 70% training, 15% validation, and 15% test data. Table I summarizes the results for the validation and test sets after training the model with 15K iterations. A qualitative example can be found in the full version 3^0.

TABLE I: Performance of the model on validation and test data and trained with 15,000 iterations.

Average Precision	Validation	Test
Overall	**0.903**	**0.901**
Abstract	0.973	0.975
Author	0.920	0.917
Email	0.870	0.859
Address	0.890	0.892
Date	0.809	0.809
Journal	0.875	0.879
Affiliation	0.888	0.876
DOI	0.950	0.945
Title	0.953	0.957

The model achieves an average precision (AP) for both the validation set ($AP \approx 90.363$) and the test set ($AP \approx 90.581$) when trained using 15,000 iterations. These AP-values validate the capability *MexPub* to accurately extract metadata from a diverse range of layouts and styles of German scientific publications. The results reveal that the model performs exceptionally well for detecting abstracts, DOIs, and titles. We hypothesize that the high precision for DOIs might be due to its pattern having a persistent format across documents; It usually starts with "*https://doi,*" followed by a chain of digits and punctuation. Regarding abstracts and titles, we assume that the high precision is partly explained by three factors; First, both patterns usually cover relatively large areas on the first pages of scientific documents, which the model better detects than smaller patterns. Specifically, our results show that for the validation set, the model achieves an $AP \approx 95.273$ for large pattern (areas $> 96^2$ pixels), while for small pattern (areas $< 32^2$ pixels), the $AP \approx 74.173$. Second, in contrast to other metadata patterns, such as emails, addresses, or affiliation names, abstracts and titles are present in almost all training data samples. Finally, the positions of title and abstract are relatively consistent across layouts, whereas those of other metadata records vary more. Furthermore, we assume *MexPub* to benefit from the source model being trained to detect titles already. However, there are cases in which the model could not detect abstracts and titles accurately. For example, when the documents do not contain an abstract, the model tends to falsely classify the first paragraph of the paper to be such.

Moreover, the results reveal that dates are the most challenging pattern for the model to detect because the cover a relatively small area on the images. We noticed that the model does not perform well on objects with an area smaller than 322 pixels than on larger ones. Furthermore, their positions on the page differ to a great extent across different layouts. Based on the assumption that the model factors in positional information, we assume that patterns with high positional variance across layouts are particularly challenging to recognize.

As a baseline approach, we compared *MexPub* against *GROBID* method [8]. To this end, we selected 100 scientific documents with different layouts from SSOAR[11] that served as inputs to both pipelines. To allow for a fair comparison, we only included papers with a layout different from those we trained *MexPub* with. Note that these samples and their layouts are not seen in the training phase of *MexPub*. Moreover, since the model was trained only on German scientific papers' first pages, we also considered only the first pages from these samples. To evaluate the models' performances, we manually parsed all metadata records on the documents and used them as ground truth patterns. Since *MexPub* outputs the coordinates of bounding boxes for each detected metadata patterns, we extract the text from the PDF documents in the areas corresponding to these coordinates.

For both methods, we created confusion matrices with the predicted values as columns and the ground truth values as rows. To compute whether the extracted metadata matched the

[7]http://labelme.csail.mit.edu/Release3.0/
[8]http://cocodataset.org/#home
[9]https://dblp.org/xml/release/
[10]https://de.wikipedia.org/wiki/Liste_der_Hochschulen_in_Deutschland
[11]https://www.gesis.org/ssoar/home

ground truth data, we used cosine similarity with a threshold of 0.85. Whenever an extracted metadata record's cosine similarity was greater than or equal to the threshold when compared with the ground truth value, we consider it as true positive. The reasons of using soft matching by allowing a dissimilarity of 0.15 is because some extracted patterns were more or less fine-grained than the ground truth. For example, the volume was not annotated as a part of the journal name, but both methods *MexPub* and *GROBID* include in in some cases while also eliminate it in other cases. Based on the confusion matrices, we computed average precision, average recall, and average F1 scores for the nine classes.

Table II depicts the results of the systems' evaluations. *MexPub* is able to correctly predict titles and authors with an F1 score of 0.940 and 0.750, respectively. However, the results reveal that it produces a significant amount of false-positives for abstracts, which is explained by the low precision value for this class.

After analysing the obtained results, it is found that *MexPub* tends to classify paragraphs that appear in the upper half of a document as abstracts, although they often represent introductions or other sections of the document's body. Similarly, the number of false-positives for journal and author is also relatively high. Furthermore, *MexPub* did not predict any of the six present patterns corresponding to affiliations. Compared to *MexPub*, *GROBID* achieves a high precision, where it produces a significantly lower amount of false-positives across all classes present in the ground truth data. However, *GROBID* fails at extracting a lot of true positive patterns which is reflected in the low recall for all classes. For some classes like "address" and "journal", *GROBID* completely fails at extracting any of the corresponding patterns.

TABLE II: Performances of GROBID and *MexPub* regarding precision, recall, and F1-score per class. NaN values represent that the class was not present in the ground-truth data (e.g. DOI), or that a model is not designed to extract the corresponding pattern (e.g. date). For a fair comparison, the macro and micro averages are computed only using the classes associated with (*).

	MexPub			GROBID		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Title*	0.934	**0.947**	**0.940**	**0.965**	0.577	0.723
Author*	0.670	**0.851**	0.750	**0.982**	0.609	**0.752**
Journal*	**0.147**	**0.385**	**0.212**	0.000	0.000	0.000
Affiliation*	0.000	0.000	0.000	**1.000**	**0.118**	**0.210**
Abstract*	0.219	**0.833**	0.346	**0.972**	0.593	**0.739**
DOI	NaN	NaN	NaN	NaN	NaN	NaN
Address*	**0.125**	**1.000**	**0.222**	0.000	0.000	0.000
Email*	0.000	0.000	0.000	**1.000**	**0.500**	**0.667**
Date	0.250	1.000	0.400	NaN	NaN	NaN
Macro average	0.299	**0.574**	0.353	**0.703**	0.342	**0.442**
Micro average	0.558	**0.754**	**0.613**	**0.766**	0.447	0.559

Although *MexPub* does not consider any contextual features, it could achieve a good results on unseen documents with completely different layouts. However, the main limitation of this model is its low generalizibility to publications with a significantly different structure. Therefore, we assume that by integrating contextual/textual features, the model can perform better.

V. Conclusion

In this paper, we proposed *MexPub* that automatically extracts metadata from scientific papers using deep learning. Contrary to conventional approaches, *MexPub* treats the PDF document as image and extracts metadata on a pixel-by-pixel level. For this, we adopted a deep learning model dedicated for object detection and retrained to detect patterns such as *text* and *figures* from PDF documents. We refine-tuned this model using the synthetic dataset proposed in this paper in order to extract fine grained patterns (e.g. title, author). The experimental results validates the capability of this approach to extract metadata from different layouts.

For future work, we will train the model on a greater variety of layouts to improve its generalizability. Future research will incorporate text-based processing in a joint neural network architecture. The visual part of this model is supposed to capture the structural characteristics of the PDF document, while the textual part is supposed to capture the semantic and contextual characteristics.

References

[1] Z. Boukhers, S. Ambhore, and S. Staab. An end-to-end approach for extracting and segmenting high-variance references from pdf documents. In *JCDL*, 2019.

[2] D. D. A. Bui, G. Del Fiol, and S. Jonnalagadda. Pdf text classification to leverage information extraction from publication reports. *Journal of biomedical informatics*, 61:141–148, 2016.

[3] G. Colavizza and M. Romanello. Citation mining of humanities journals: The progress to date and the challenges ahead. *Journal of European Periodical Studies*, 4(1):36–53, 2019.

[4] H. Kaiming, G. Georgia, D. Piotr, and G. Ross. Mask r-cnn. page 2961–2969. IEEE international conference on computer vision.

[5] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[6] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *JCDL*, page 385. ACM Press, 2013.

[7] R. Liu, L. Gao, D. An, Z. Jiang, and Z. Tang. Automatic document metadata extraction based on deep networks. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 305–317. Springer, 2017.

[8] L. Romary and P. Lopez. Grobid-information extraction from scientific publications. *ERCIM News*, 2015.

[9] N. Siegel, N. Lourie, R. Power, and W. Ammar. Extracting Scientific Figures with Distantly Supervised Neural Networks. *JCDL*, pages 223–232, 2018.

[10] J. Stadermann, S. Symons, and I. Thon. Extracting hierarchical data points and tables from scanned contracts. *UIMA@ GSCL*, 1038, 2013.

[11] C. Stahl, S. Young, D. Herrmannova, R. Patton, and J. Wells. DeepPDF: A Deep Learning Approach to Analyzing PDFs. 2018.

[12] A. Takasu. Bibliographic attribute extraction from erroneous references based on a statistical model. In *ACM/IEEE-CS joint conference on Digital libraries*, pages 49–60. IEEE Computer Society, 2003.

[13] D. Tkaczyk. New Methods for Metadata Extraction from Scientific Literature. *CoRR*, abs/1710.10201, 2017.

[14] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016.

[15] H. t. Yang. Pipelines for Procedural Information Extraction from Scientific Literature: Towards Recipes using Machine Learning and Data Science. In *ICDAR Workshop*, pages 41–46, 2019.

[16] X. Zhong, J. Tang, and A. J. Yepes. Publaynet: largest dataset ever for document layout analysis. In *ICDAR*, 2019.