

DocParser: Hierarchical Document Structure Parsing from Renderings

Johannes Rausch,¹ Octavio Martinez,¹ Fabian Bissig,¹
Ce Zhang,¹ Stefan Feuerriegel²

¹ Department of Computer Science, ETH Zurich

² Department of Management, Technology, and Economics, ETH Zurich
johannes.rausch@inf.ethz.ch, octaviom@student.ethz.ch, fbissig@student.ethz.ch
ce.zhang@inf.ethz.ch, sfeuerriegel@ethz.ch

Abstract

Translating renderings (e.g. PDFs, scans) into hierarchical document structures is extensively demanded in the daily routines of many real-world applications. However, a holistic, principled approach to inferring the complete hierarchical structure of documents is missing. As a remedy, we developed “DocParser”: an end-to-end system for parsing the complete document structure – including all text elements, nested figures, tables, and table cell structures. Our second contribution is to provide a dataset for evaluating hierarchical document structure parsing. Our third contribution is to propose a scalable learning framework for settings where domain-specific data are scarce, which we address by a novel approach to weak supervision that significantly improves the document structure parsing performance. Our experiments confirm the effectiveness of our proposed weak supervision: Compared to the baseline without weak supervision, it improves the mean average precision for detecting document entities by 39.1 % and improves the F1 score of classifying hierarchical relations by 35.8 %.

1 Introduction

The structural and layout information in a document can be a rich source of information that facilitates Natural Language Processing (NLP) tasks (e.g. information extraction). Over the years, the NLP community has developed a range of techniques to *detect*, *understand*, and *take advantage of* document structures (Hurst and Nasukawa 2000; Chen, Tsai, and Tsai 2000; Tengli, Yang, and Ma 2004; Luong, Nguyen, and Kan 2012; Govindaraju, Zhang, and Ré 2013; Katti et al. 2018; Schäfer et al. 2011; Schäfer and Weitz 2012; Gancarek et al. 2020).

However, structural information in documents is becoming increasingly challenging to obtain — many file formats that are prevalent today are being rendered without structural information. Prominent examples are PDF documents: this file format benefits from portability and immutability, yet it is flat in the sense that it stores all content as isolated entities (e.g., combinations of characters and positions) and, thus, hierarchical information is lacking. As such, the structure behind figures and especially tables is discarded and thus no longer available to computerized analyses in NLP.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In contrast, file formats such as XML or JSON naturally encode hierarchical document structures among textual entities. Hence, techniques are required in order to convert renderings into structured, textual document representations to enable *joint inference* between text, layout, and other document structures.

Earlier attempts for structure parsing on documents focused on a subset of simpler tasks such as segmentation of text regions (Antonacopoulos et al. 2009), locating tables (Zanibbi, Blostein, and Cordy 2004; Embley et al. 2006), or parsing them (Schreiber et al. 2018), but not parsing complete document structures. However, document structures are required as a representation of many downstream tasks in NLP. For instance, recent efforts in the NLP community (Katti et al. 2018; Apostolova and Tomuro 2014; Liu et al. 2019) have shown that utilizing 2D document information, e.g. character and word positions, can be an effective way to improve upon standard NLP tasks such as information extraction.

A holistic, principled approach for inferring the complete hierarchical structure from documents is missing. On the one hand, such a task is nontrivial due to the complexity of documents, particularly their deeply-nested structures. For instance, nested tables are fairly easy to recognize for human readers, yet detecting them is known to impose computational hurdles (cf. Schreiber et al. 2018). On the other hand, efficient learning is prevented as large-scale training sets are lacking (cf. Arif and Shafait 2018; Schreiber et al. 2018). Notably, prior datasets are limited to table structures (Gobel et al. 2013; Rice, Jenkins, and Nartker 1995) and not the complete document structures. Needless to say, complex structures also make the labeling process significantly more costly (Wang, Phillips, and Haralick 2004). Therefore, an effective implementation that makes only a scarce use of labeled data is demanded.

This work focuses on parsing the hierarchical document structure from renderings. We develop an end-to-end system for inferring the complete document structure (see Figure 1). This includes all entities (e.g., text, bibliography regions, figures, equations, headings, tables, and table cells), as well as the hierarchical relations among them. We specifically adapt to settings in practice that suffer from data scarcity. For this purpose, we propose a novel learning framework for scalable weak supervision. It is intentionally tailored to the

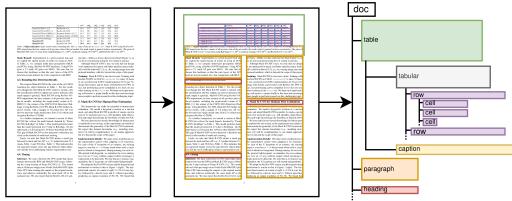


Figure 1: DocParser takes rendered document images (left) as input, performs segmentation into bounding boxes (center), and then outputs the hierarchical structure of the full document (right). Shown is an illustrative sketch; examples are provided in the supplements.

specific needs of parsing document renderings; that is, we create weakly-supervised labels by utilizing the reverse rendering process of L^AT_EX. The reverse rendering returns the bounding boxes of all entities in documents together with their category (e.g., whether the entity is a table or a figure, etc.). Yet the outcomes are noisy (i.e., imprecise bounding boxes, missing entities, incorrect labels) and without deep structure information (e.g. information such as table row numbers is missing). Nevertheless, as we shall see later, the generated data greatly facilitates learning by being treated as weak labels.

Contributions:¹ We extend prior literature on document parsing in the following directions:

1. We contribute “DocParser”. This presents the first end-to-end system for parsing renderings into *hierarchical* document structures. Prior literature has merely focused on simpler tasks such as table detection or table parsing but not on the parsing of complete documents. As a remedy, we present a system for inferring document structures in a holistic, principled manner.
2. We contribute the first dataset (called “arXivdocs”) for evaluating document parsing. It extends existing datasets for parsing in two directions: (i) it includes all entities that can appear in documents (i.e. not just tables) and (ii) it includes the hierarchical relations among them. The dataset is based on 127,472 scientific articles from the arXiv repository.
3. We propose a novel weakly-supervised learning framework to foster efficient learning in practice where annotated documents are scarce. It is based on an automated and thus scalable labeling process, where annotations are retrieved by reverse rendering the source code of documents. Specifically, in our work, we utilize T_EX source files from arXiv together with synctex for this objective. This then yields weakly-supervised labels by reverse rendering of the T_EX code.
4. We conduct extensive evaluation of our proposed techniques, outperforming the state-of-the-art on the related task of table parsing.

¹Source codes and the arXivdocs dataset are available from <https://github.com/DS3Lab/DocParser>.

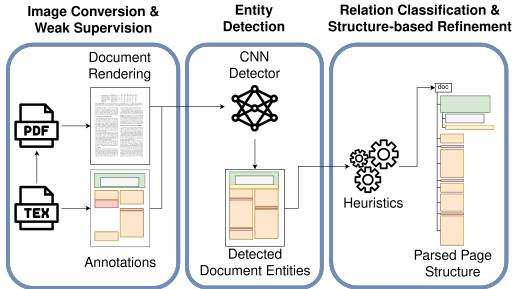


Figure 2: System overview.

2 DocParser System

2.1 Problem Description

Given a set of document renderings D_1, \dots, D_n , the objective is to generate hierarchical structures T_1, \dots, T_n . A hierarchical structure T_i , $i = 1, \dots, n$, consists of both entities and relations as follows:²

Entities E_j , $j = 1, \dots, m$, refer to the various elements within a document, such as a figure, table, row, cell, etc. Each entity is described by three attributes: (1) its semantic category $c_j \in \mathcal{C} = \{C_1, \dots, C_l\}$ (i.e., which defines the underlying type) and (2) the coordinates given by rectangular bounding box B_j in the document rendering. There further is (3) a confidence score P_j . This is not part of the ground truth labels; however, it comes from the predictions inside the DocParser system.

Relations R_j , $j = 1, \dots, k$ of type Ψ are given by triples $(E_{\text{subj}}, E_{\text{obj}}, \Psi)$ consisting of a subject E_{subj} , an object E_{obj} , and a relation type $\Psi \in \{\text{parent_of}, \text{followed_by}, \text{null}\}$. The latter, null, is reserved for entities with meta-information that do not have designated order (i.e., header, footer, keywords, date, page number). All other entities must have $\Psi \neq \text{null}$.

The combination of entities and relations is sufficient to reconstruct the hierarchical structure T_i for a document. However, generating such a hierarchical structure from a document rendering D_i is subject to inherent challenges: the similar appearance of entities impedes detection and, further, the hierarchy can be nested arbitrarily, with substantial variation across different documents.

2.2 System Components

DocParser performs document structure parsing via 5 components (see overview in Figure 2): (1) image conversion, (2) entity detection, (3) relation classification, (4) structure-based refinement, and (5) scalable weak supervision. To store document structures, we developed a customized, JSON-based file format.

Component 1: Image Conversion

²For consistency, we use the term “entity” throughout the article when referring to all elements in the document structure (e.g. a figure, table, or text) that need to be detected. While the term “object” is common in computer vision, we chose the term “entity” to highlight its semantic nature for NLP.

Document renderings are converted into images with a predefined resolution ρ . Furthermore, all images are resized to a fixed rectangular size ϕ (if necessary, with zero padding).

The document images are further pre-processed: the RGB channels of all document images are normalized analogous to the MS COCO dataset (i. e., by subtracting the mean RGB channel values from the inputs). The reason is that all neural models are later initialized with pre-trained weights from the MS COCO dataset (Lin et al. 2014).

Component 2: Entity Detection

To detect all document entities within a document image, we build upon a neural model for image segmentation, namely Mask R-CNN (He et al. 2017). Specifically, it takes the images from the previous component as input and then returns a flat list of entities E_1, \dots, E_m as output. For each entity Mask R-CNN determines (i) its rectangular bounding box, (ii) confidence score, (iii) a binary segmentation mask that distinguishes between the detected entity and background pixels within the bounding box, and (iv) a category label for the entity. Our implementation makes use of 23 categories C : CONTENT BLOCK, TABLE, TABLE ROW, TABLE COLUMN, TABLE CELL, TABULAR, FIGURE, HEADING, ABSTRACT, EQUATION, ITEMIZE, ITEM, BIBLIOGRAPHY BLOCK, TABLE CAPTION, FIGURE GRAPHIC, FIGURE CAPTION, HEADER, FOOTER, PAGE NUMBER, DATE, KEYWORDS, AUTHOR, AFFILIATION.³

Component 3: Relation Classification

A set of heuristics is applied to translate the flat list of entities into hierarchical relations R_1, \dots, R_k . Here, we distinguish the heuristics according to whether they generate (1) the **nesting** among entities or (2) the **ordering** for entities of the same nesting level. The former case corresponds to $\Psi = \text{parent_of}$, while the latter determines all relations with $\Psi = \text{followed_by}$. In this component, we ignore all entities with meta-information, e. g. footers, as these have no designated hierarchy (cf. document grammar in the supplements).

Relations with Nesting (parent_of): Four heuristics h_1, \dots, h_4 determine parent-child relation as follows:

(h_1 : Overlaps) A list of candidate parent-child relations is compiled based on the overlap of bounding boxes. That is, DocParser loops over all bounding boxes and, for each bounding box B_{subj} , it determines all other bounding boxes that are contained within B_{subj} .

Formally, this is given by all tuples of bounding boxes $(B_{\text{subj}}, B_{\text{obj}})$ with $\text{subj} \in m$, $\text{obj} \in m$, and $\text{subj} \neq \text{obj}$ where $h_1(B_{\text{subj}}, B_{\text{obj}})$ is satisfied: Tuples for which the bounding box of B_{obj} is fully or partially enclosed by the bounding box of B_{subj} are added to the candidate list. Furthermore, we add tuples to the candidate list that satisfy $\frac{\text{area}(B_{\text{subj}} \cap B_{\text{obj}})}{\text{area}(B_{\text{obj}})} \geq \theta_1$ and $\frac{\text{area}(B_{\text{subj}})}{\text{area}(B_{\text{obj}})} > \theta_2$, i. e. they must have a certain overlap fraction θ_1 and size ratio θ_2 . In DocParser, thresholds of $\theta_1 = 0.45$ and $\theta_2 = 1.2$ are used.

(h_2 : Grammar Check) This heuristic validates the candidate list against a predefined document grammar (see doc-

ument grammar in the supplements). Concretely, all illegal candidates, e.g., a TABULAR nested inside a FIGURE, are removed.

(h_3 : Direct children) The candidate list is further pruned so that it contains only direct children of the parent and not sub-children. For this purpose, all sub-children are removed. As an example, this should remove $(E_{\text{subj}}^1, E_{\text{obj}}^3)$ from a candidate list $\{(E_{\text{subj}}^1, E_{\text{obj}}^2), (E_{\text{subj}}^1, E_{\text{obj}}^3), (E_{\text{subj}}^2, E_{\text{obj}}^3)\}$, since it represents a sub-child and not a direct child of E_{subj} .

(h_4 : Unique Parents) The candidate list is altered so that each entity has only a single parent. Formally, if an entity E_{obj} has multiple candidate parents, we first compare the Intersection over Union (IoU) of the bounding boxes of all candidate parents with E_{obj} : $\text{IoU} = \frac{\text{area}(B_{\text{subj}} \cap B_{\text{obj}})}{\text{area}(B_{\text{subj}} \cup B_{\text{obj}})}$. We then keep the parent with the maximal IoU, while all others are removed. If two parents have the same IoU, we select the element with the highest confidence score P_j as parent. If that value is also equal, we choose the entity with the largest bounding box.

Relations with Ordering (followed_by): The entities are ordered according to the general reading flow (i. e., from left to right). Here care is needed so that multi-column pages are processed correctly. For this, two heuristics o_1 and o_2 are used. By default, all entities are processed by both heuristics. Children of floating entities are only processed by heuristic o_2 , however.

(o_1 : Page Layout Entities) First, all entities are grouped according to their coordinates on the document page, namely, into groups belonging to the (a) left side G_l , (b) center G_c , or (c) right side G_r . Formally, this is achieved by computing the overlap for each entity E_j , $j = 1, \dots, m$ with the left (and right) side of a document page, i. e., $\tau_{\text{ovlp}} = \text{overlap}/\text{width}(B)$. If the overlap with either the left (or the right) side is above a threshold (i. e., $\tau_{\text{ovlp}} > 0.7$), the entity E_j is assigned to the left (or right) side.

Otherwise, if such assignment is not possible with high confidence, the entity E_j is assigned to center group G_c . In essence, the center group is an indicator whether the document is in single- or multi-column.

If no entities have been assigned to the center group (i. e., $G_c = \emptyset$), then the entities are ordered first according to G_l followed by G_r . Within each group, the entities are ordered top-to-bottom and then left-to-right by applying heuristic o_2 . In sum, this approach should find an appropriate ordering for multi-column pages. If entities have been assigned to the center group (i. e., $G_c \neq \emptyset$), then grouping is further decomposed into additional subgroups: the entities $E \in G_c$ from the center group are used to split G_l , G_c , and G_r into vertical subgroups G_l^ℓ , G_c^ℓ , and G_r^ℓ , respectively. Afterward, we loop over all vertical subgroups ℓ . For each, we order the entities according to the group (first G_l^ℓ , followed by G_c^ℓ and then G_r^ℓ). Within each subgroup, we perform the ordering via heuristic o_2 . This approach should correctly arrange entities in two cases: (1) in single-column pages and (2) when multi-column pages are split into different chunks by full-width figures or tables.

For each subgroup, we perform the ordering via heuristic o_2 .

³For consistency, this formatting is utilized for all entities.

(o₂: Reading Flow) The entities E_j , $j = 1, \dots, m$, are ordered top-to-bottom and, within lines, left-to-right, so that it matches the usual reading flow in documents. Formally, let the top-left corner of a document image refer to the coordinate $(0, 0)$. Furthermore, let us consider the top-left location of all bounding boxes B_j . The top-left location is then used to sort the entities first by their y -coordinate of B_j and, if equal, by their x -coordinate (both ascending).

Component 4: Structure-Based Refinement We utilize the classified relations to iteratively refine entities and relations in four steps when parsing full document pages:

(1) For each entity E_{parent} with l child entities $E_{\text{child}}^1, \dots, E_{\text{child}}^l$, we update its bounding box such that $B_{\text{parent}} = \text{union}(B_{\text{parent}}, B_{\text{child}}^1, \dots, B_{\text{child}}^l)$. (2) For parent entities E_{parent} with exactly one child entity of the same category, we remove the child entity and update B_{parent} such that it is the union of parent and child bounding boxes. We also consider entity pairs of categories that do not conform to the document grammar. This allows us to dismiss duplicate entities of any category. (3) If an entity E_{child} is sibling to other entities in a way that conflicts the document grammar, we generate a new entity that encloses E_{child} to achieve conformity with the document grammar. Concretely, nested FIGURE structures are defined such that one FIGURE should at most contain one FIGURE GRAPHIC entity child. If multiple FIGURE GRAPHIC are classified as children, we wrap each of them individually into new FIGURE entities. (4) If no parent is found for an entity E_{child} that should only occur as a child entity, we identify a suitable parent entity by analyzing its neighboring siblings as follows: we consider all entities that jointly appear in an ordering relation with E_{child} as a candidates E_{cand} . We dismiss candidates of category \mathcal{C} that would not conform to the hierarchies defined in the document grammar. Finally, we dismiss any candidate for which $B_{\text{cand}} \cap B_{\text{child}} = \emptyset$. If exactly one candidate remains, we update its bounding box $B_{\text{cand}} = \text{union}(B_{\text{cand}}, B_{\text{child}})$.

The updates to the set of entities can lead to further changes to the classified relations. For this reason, whenever changes are made to entities in one of the four refinement steps, we update the relations via Component 3 and move back to refinement step (1). The refinement is completed once no change is applied in any of the steps or a limit of r loop iterations has been reached.⁴

Component 5: Scalable Weak Supervision

The system is further extended by scalable weak supervision. This aims at improving the performance of entity detection and, as a consequence, of end-to-end parsing.

Our weak supervision builds upon an additional dataset that consists of source codes (rather than document renderings). The source codes allow us to create a mapping between entities in the source code and their renderings. This process has three particular characteristics: first, the mapping is noisy and thus creates only weak labels. Despite that, the weak labels can aid efficient learning. Second, annotations are obtained only for some entities and relations. Third, if automated, this process circumvents human annotations

⁴Details on our parameter choice and pseudocode are included in the supplements.

and is thus highly scalable.

Let the unlabeled entities found in the source code be given by S_1, \dots, S_k . For them, we generate weak labels W_1, \dots, W_k consisting of a semantic category and coordinates of the bounding box. However, both the semantic category and the bounding box can be subject to noise. Furthermore, weak labels are generated merely for a subset $\mathcal{C}' \subseteq \mathcal{C}$ of the semantic categories.

In DocParser, the weak supervision is based on \TeX source files that are used to generate document renderings in the form of PDF files. The mapping between both formats is then obtained via `synctex` (Laurens 2008). `synctex` is a synchronization tool that performs a reverse rendering, so that PDF locations are mapped to \TeX code. For given coordinates in the document rendering, `synctex` returns a list of rectangular bounding boxes and the corresponding source code. Notably, the inference bounding boxes represent noisy labels, since the resulting entity annotations could be wrongly labeled, shifted, or entirely missing.

We proceed as follows. We iterate through the source code and retrieve bounding boxes for all \TeX commands. We then map the source code to our entities E . For instance, the bounding box for \TeX code `\includegraphics` inside a `\begin{figure} ... \end{figure}` environment is mapped onto a FIGURE_GRAPHIC entity that is nested inside a FIGURE entity. Bounding boxes for all entities that act as inner children are created dynamically by computing the union bounding of all child bounding boxes.

We perform following processing steps to generate noisy labels for weak supervision:

1. Bounding boxes that are retrieved for simple text tokens inside the source code are mapped to CONTENT LINE entities.
2. If we encounter environments or commands (e.g., `\begin{itemize}` or `\item`), we create corresponding candidate entities. All entities retrieved for tokens inside the scope of these environments are created as nested child entities. This approach is used to create the following entity types, namely FIGURE, FIGURE GRAPHIC, FIGURE CAPTION, TABLE, TABULAR, TABLE CAPTION, ITEMIZE, ITEM, ABSTRACT, and BIBLIOGRAPHY. Any other entities are mapped onto the CONTENT LINE category.
3. We utilize a special characteristic of `synctex` to identify EQUATION, EQUATION FORMULA and EQUATION LABEL entities: bounding boxes returned by `synctex` are highly uniform and typically consist of per-line bounding boxes of consistent width and x -coordinates. Equations and labels are an exception to this rule and typically only consist of vertically aligned bounding boxes of smaller width.
4. The sectioning structure of documents is considered: any type of `section` command is mapped to a SECTION entity. The argument of the sectioning command, e.g. `\subsection{titlearg}` is mapped via `synctex` to a HEADER entity. Entities generated from code in the scope of a section are created as children to the section entity that corresponds to the current section scope.

5. Within sections, we sort entities based on a top-to-bottom, left-to-right reading order. Using these sorted lists of sibling entities, we form CONTENT BLOCK entities from subsequent groups of CONTENT LINE entities within page columns. If such block occurs within a BIBLIOGRAPHY environment, we instead map it to a BIBLIOGRAPHY BLOCK entity.
6. In TABLE environments, we consider all child entities (except captions) that do not span across a whole table width as CELL and the remainder as TABLE ROW. As we shall see later, this is effective at retrieving complex table structures.
7. We use the detected table cells to generate rows and columns as follows: We compute the centroids of all cells. To identify rows, we consider the sorted y -coordinates of the centroids and group them such that the pixel-wise distance between two consecutive y -coordinates in a group is smaller or equal to 5. If any identified group contains two or more centroid y -coordinates, we create a TABLE ROW entity from the union of the corresponding TABLE CELL entities. Analogously, using the x -coordinates of the cell centroids, we identify TABLE COLUMN entities.
8. Additional cleaning steps are performed for tables and figures: Child entities with width or height of 2 or fewer pixels are discarded. Caption bounding boxes that enclose other non-caption child entities are also discarded.
9. We make sure that entities contain at most one leaf node by moving excess leaves into newly generated CONTENT LINE entities.
10. We remove duplicate bounding boxes and entities without any leaf nodes in their respective sub-tree. Candidates are filtered such that only a group of entities and their respective sub-tree are preserved: ITEMIZE, FIGURE, TABLE, EQUATION, HEADING, CONTENT BLOCK, BIBLIOGRAPHY, ABSTRACT.

During training, entities with obvious errors are dismissed, i. e. leaf nodes or entities with bounding boxes that extend beyond page limits or with area of 0.

3 Datasets with Document Structures

We contribute the dataset ‘‘arXivdocs’’ that is tailored to the task of hierarchical structure parsing. It comes in two variants: **arXivdocs-target** and **arXivdocs-weak**. (1) arXivdocs-target contains documents that have been manually checked and annotated. (2) arXivdocs-weak contains a large-scale set of documents that have no manual annotations but that can be used for weak supervision.

3.1 arXivdocs-target

arXivdocs-target provides a set of documents with manual annotations of the complete document structure. These documents were randomly selected from arXiv as an open repository of scientific articles, but in a way such that each has at most 30 pages and contains at least one TABLE within the source code. Altogether, it counts 362 documents. arXivdocs-target comes with predefined splits for *training*,

validation, and *eval* that consist of 160, 79, 123 documents, respectively. The dataset comprises of 30 different entity categories.⁵ We ensure a fairly uniform distribution of entity categories across different splits by sampling one random page rendering for each of the 362 documents that contain an ABSTRACT, FIGURE, or TABLE. On average, each document contains 86.32 entities. The number of leaf nodes in the document graph as well as the frequency and average depth of the different entities are reported in the supplements.

Evidently, the most common category in the dataset is CONTENT LINE (34.33 %). This is because they typically represent leaf nodes in the graph and are children of larger entities such as ABSTRACT, CAPTION, or CONTENT BLOCK.

Annotators were instructed to follow the document grammar during labeling. Annotation of disallowed hierarchies is, however, possible to provide them the freedom to deal with the range of different document representations. Document annotations are automatically initialized by our scalable weak supervision mechanism to speed up the annotation process. The labelers were instructed to annotate entities only up to the coarseness that is used by DocParser, e. g. labeling content blocks, rather than individual lines.

3.2 arXivdocs-weak

arXivdocs-weak contains 127,472 documents with an average length of 12.84 pages that were retrieved from arXiv. We selected only documents that have a length of at most 30 pages and contain at least one TABLE within their source code. For reproducibility, we make our weak labels available.⁶

4 Computational Setup

4.1 Mask R-CNN

Mask R-CNN extends the architecture of a convolution neural network with skip connections (He et al. 2016) so that it is highly effective for image segmentation and entity detection.⁷ Formally, it comprises of multiple stages with decreasing spatial resolution. The output of these stages is then fed into a so-called feature pyramid network (FPN) (Lin et al. 2017). The FPN then interconnects these inputs in multiple stages of increasing spatial resolution to produce multi-scale feature maps. Specifically, we use a ResNet-110 architecture (He et al. 2016) to extract features in 5 stages at different resolutions. The outputs of stages 2 to 5, denoted as C_2, \dots, C_5 , are passed to the FPN. The FPN outputs a total of 5 feature maps P_2, \dots, P_6 at different resolutions. We refer the reader to (Lin et al. 2017) for a detailed description of the five feature maps. The multi-scale feature maps

⁵Some entity categories are extremely rare and, hence, only a subset is later used as part of our experiments.

⁶For this purpose, the dataset was labeled via our proposed weak supervision mechanism and thus contains both entities E_j and hierarchical relations R_j . For reasons of space of the physical files, bounding boxes are only stored for entities in leaf nodes. For all other entities, the bounding boxes can be calculated by taking the union bounding box of their children.

⁷A model illustration is included in the supplements.

are then input to different prediction networks: first, a region proposal network (RPN) generates a list of candidate bounding boxes that should contain an entity. Second, a Region of Interest (RoI) alignment layer filters out the multi-scale feature maps that correspond to the candidate regions. We note that all 5 feature maps are used by the RPN, but P_6 is not included in the inputs to the RoI alignment layer. Third, for each region proposal, a mask sub-network predicts the segmentation masks, based on the RoI aligned features. These segmentation masks are not used in subsequent steps of DocParser at prediction time; however, they are utilized in our loss function during the training process. Fourth, these bounding boxes are subsequently refined in a detection sub-network, thereby yielding the final bounding boxes B . It also provides the label for classifying the entity category.

All of the above sub-networks were carefully adapted to the specific characteristics of our task: (1) We modified the region proposal network so that it uses a maximum base aspect ratio of 1:8 per entity. The reason for this modification is that document entities (as opposed to classical image segmentation) contain entities that have highly rectangular shapes. This is the case for most entities, e.g., single CONTENT LINE or TABLE ROW entities. (2) The output size of the classifier sub-network is modified so that it can produce predictions for entities across all semantic categories \mathcal{C} . (3) During training of the mask sub-network, we treat all pixels in ground truth bounding boxes as foreground. We do this to incorporate our understanding of the exact shape of many entities that span very wide rectangular regions. (4) We use a mask sub-network loss with a weighting factor of 0.5. This is to prioritize that features relevant for the correct prediction of bounding boxes and entity categories are learned. The Mask R-CNN stage of DocParser comprises 63,891,032 parameters and is built upon the implementation of Mask R-CNN provided by Abdulla (2017), yet which we carefully adapted as described above.

Training Procedure: All neural models are initialized with pre-trained weights based on the MS COCO dataset (Lin et al. 2014). We then train each model across three phases for a total of 80,000 iterations. This is split into three phases of 20,000, 40,000, and 20,000 iterations, respectively. During the first phase, we freeze all layers of the CNN that is used as the initial block in Mask R-CNN. In the second phase, stages four and five of the CNN are unfrozen. In the last phase, all network layers are trainable. Early stopping is applied based on the performance on the validation set for unrefined predictions. The performance is measured every 2000 iterations via the so-called intersection over union with a threshold of 0.8.

We train all models in a multi-GPU setting, using 8 GPUs with a vRAM of 12 GB. Each GPU was fed with one image per training iteration. Accordingly, the batch size per training iteration is set to 8. Furthermore, we use stochastic gradient descent with a learning rate of 0.001 and learning momentum of 0.9.

Parameter Settings: During training, we sampled randomly 100 entities from the ground truth per document image (i.e., up to 100 entities as some document images might have fewer). In Mask R-CNN, the maximum number of entity

predictions per image is set to 200. During prediction, we only keep entities with a confidence score P_j of 0.7 or higher.

Weak Supervision: Training with weak supervision is as follows: all models are initialized with the weights of our pre-trained DocParser WS instead of default weights. We perform the training with learnable parameters analogous to phase 1 above but for 2000 steps with early stopping. In our experiments, we use only a subset of 80 % of the annotated documents from arXivdocs-weak, while the other 20 % remain unused. The intention is that we want to allow for additional annotations in the future while ensuring comparability to our results. We further ensure a fairly uniform distribution of entities by utilizing only document pages that contain at least an ABSTRACT, a FIGURE, or TABLE, while all others are discarded. This amounts to 593,583 pages.

4.2 System Variants

We compare the following variants of DocParser: **DocParser Baseline** is trained solely on the noise-free labels provided for the training dataset (here: arXivdocs-target); **DocParser WS** benefits from weak supervision (WS). It is trained based on a second dataset (here: arXivdocs-weak) with noisy labels for weak supervision. This is to test whether training systems on noisy labels can lead to higher performance, compared to training on small but noise-free training datasets; **DocParser WS+FT** is initialized with the weights from DocParser WS, but then fine-tuned (FT) on the target dataset.

4.3 Performance Metrics

We separately evaluate the performance of our system for (i) detection of entities E_j and (ii) classification of hierarchical relations R_j . The former aims at a high detection rate (i.e. recognizing true positives out of all positives). Hence, we use the average precision as evaluation metric. The latter is based on the F1 score as it represents a typical classification task (i.e. recognizing one of the relations from Ψ).

Entity Detection: entity detection is commonly measured by the mean average precision (mAP) of a model (0: worst, 100: best). The inferred entities $E_j = (c_j, B_j, P_j)$ are compared against the ground truth label consisting of the true category \hat{c}_j with a bounding box \hat{B}_j . Here we follow common practice in computer vision (Everingham et al. 2010) and measure the overlap between bounding boxes from the same category. Specifically, we calculate the so-called intersection over union (IoU): $\text{IoU} = \frac{\text{area}(B_j \cap \hat{B}_j)}{\text{area}(B_j \cup \hat{B}_j)}$. If the IoU is higher than a user-defined threshold, a predicted entity is considered a true positive. If multiple entities are matched with the same ground truth entity, we only consider the entity with the highest IoU as a true positive. Unmatched predictions and ground truth entities are considered false positives and false negatives, respectively. This is then used to calculate the average precision (AP) per semantic category $C_k \in \mathcal{C}$. The overall performance across all categories is given by the mean average precision. We compare IoU

thresholds of 0.5 and 0.65.⁸

Prediction of Hierarchical Relations: Here we measure the classification performance for predicting the correct relations. A relation $R = (E_{\text{subj}}, E_{\text{obj}}, \Psi)$ is counted as correct only if the complete tuple is identical. However, the performance depends on the correct entity detection as input. Hence, we later vary the IoU thresholds for entity detection analogous to above and then report the corresponding F1 score for correctly predicting hierarchical relations. The F1 score is the harmonic average of precision and recall for predicting these triples (0: worst, 1: best).

Note that our performance measure is relatively strict. We show that, even if some F1 scores are in a lower range, we can recover the overall document structure successfully. In particular, we outperform state-of-the-art OCR results, as illustrated in the qualitative samples in our supplements.

4.4 Robustness Check: Table Structure Parsing

We additionally train our model for structure parsing so that it identifies table structures to demonstrate the robustness of our system and weak supervision.

We confirm the effectiveness of our weak supervision as follows: we draw upon the ICDAR 2013 dataset (Gobel et al. 2013) for table structure parsing and compare it with the state-of-the-art. The ICDAR 2013 dataset consists of a variety of real-world documents and is not limited to scientific articles. We proceed analogously to full document structure parsing and train the three system variants for the task of table structure recognition.

DocParser Baseline is trained solely on the samples provided in the ICDAR 2013 training dataset; **DocParser WS** is trained on table structures generated from arXivdocs-weak. **DocParser WS+FT** is generated by subsequent fine-tuning on the ICDAR training split.⁹

Both training and fine-tuning of all variants follow the 3 phase training scheme for a total of 80,000 iterations.¹⁰

5 Results

The key focus of our experiments is to confirm the effectiveness of DocParser for parsing the complete document structures. However, we emphasize again that both suitable baselines and datasets for this task are hitherto lacking. Hence, we proceed two-fold. On the one hand, we evaluate the performance based on arXivdocs as the first dataset for document structure parsing. On the other hand, we draw upon the table structure ICDAR 2013 dataset: it is limited to table structures and not complete holistic parsing of document structures. However, it allows to test the effectiveness of our weak supervision against state-of-the-art.

⁸Additional results for IoU=0.8 are in the supplements.

⁹Details about the setting and additional experiments are provided in the supplements.

¹⁰Due to the different domain of the target dataset, we experimented with other weak supervision strategies, e.g. randomly sampling images from arXivdocs-weak and ICDAR 2013 during the same training procedure. However, the performance of models trained by sequential fine-tuning could not be surpassed.

| AP | IoU=0.5 | | | IoU=0.65 | | |
|--------------|-------------|-------------|--------------|-------------|-------------|-------------|
| | Baseline | WS | WS+FT | Baseline | WS | WS+FT |
| mean AP | 49.9 | 34.6 | 69.4 | 38.5 | 32.4 | 56.5 |
| ABSTRACT | 95.2 | 90.5 | 95.2 | 90.5 | 81.0 | 95.2 |
| AFFILIATION | 51.6 | 0.0 | 46.0 | 5.9 | 0.0 | 16.2 |
| AUTHOR | 18.0 | 0.0 | 23.6 | 20.4 | 0.0 | 16.7 |
| BIB. BLOCK | 42.4 | 79.1 | 94.7 | 43.2 | 93.9 | 80.3 |
| CONT. BLOCK | 89.3 | 69.8 | 88.4 | 83.2 | 67.0 | 84.4 |
| DATE | 0.0 | 0.0 | 24.1 | 0.0 | 0.0 | 9.3 |
| EQUATION | 65.8 | 54.5 | 82.1 | 40.6 | 52.1 | 72.8 |
| FIG. CAPTION | 47.8 | 30.5 | 69.2 | 44.0 | 17.7 | 59.5 |
| FIG. GRAPHIC | 22.3 | 5.2 | 60.2 | 15.9 | 4.4 | 54.5 |
| FIGURE | 47.8 | 35.3 | 63.5 | 44.0 | 33.9 | 59.4 |
| FOOTER | 55.7 | 0.0 | 69.3 | 48.9 | 0.0 | 59.7 |
| HEADER | 79.7 | 0.0 | 88.3 | 64.8 | 0.0 | 56.6 |
| HEADING | 53.7 | 52.1 | 66.4 | 33.1 | 46.0 | 45.4 |
| ITEM | 0.0 | 33.6 | 50.5 | 0.0 | 35.3 | 33.5 |
| ITEMIZE | 0.0 | 25.0 | 58.3 | 0.0 | 25.0 | 50.0 |
| KEYWORDS | 36.4 | 0.0 | 59.0 | 36.4 | 0.0 | 43.0 |
| PAGE NR. | 74.7 | 0.0 | 77.3 | 28.5 | 0.0 | 42.0 |
| TAB. CAPTION | 55.2 | 69.1 | 76.6 | 40.2 | 61.6 | 63.4 |
| TABLE | 84.5 | 96.3 | 94.3 | 62.7 | 87.9 | 89.6 |
| TABULAR | 78.4 | 50.8 | 100.0 | 68.4 | 42.4 | 99.5 |

Table 1: Average precision (AP) of entity detection.

5.1 Document Structure Parsing

We compare the performance of document structure parsing based on our arXivdocs-target dataset across both performance metrics.

Entity Detection The overall performance for entity detection is detailed in Table 1 (first row). We discuss the performance for $\text{IoU} = 0.5$ in the following. DocParser Baseline achieves an mAP of 49.9. This is higher than DocParser WS with an mAP of 34.6. We attribute this to the fact that several entity categories from arXivdocs-target are not part of arXivdocs-weak. Notably, the fine-tuned system DocParser WS+FT results in significant performance improvements: it obtains a mAP of 69.4, which, in comparison to the baseline DocParser, is an improvement by 39.1%.

DocParser WS+FT consistently outperforms the baseline system, even for categories that are not annotated during weak supervision (e.g. AUTHOR, FOOTER, HEADER, PAGE NUMBER). We attribute this to the better model initialization due to the prior weakly supervised pre-training. There is a small number of entity categories for which the Baseline achieves higher AP values. We attribute this to our experimental protocol which yields the best model via early stopping, based on mAP and not on individual entity AP values. For a few entities a decrease can be observed after fine-tuning (e.g. TABLE at $\text{IoU}=0.5$). We attribute this to the high quality of weak annotations for this category and, consequently, a slight decrease of generalization due to fine-tuning. Some AP values (for both DocParser Baseline and DocParser WS) amount to 0.0, e.g. for DATE. This is caused by the absence of some categories in arXivdocs-weak in the case of DocParser WS. For DocParser Baseline, we attribute this to the limited amount of samples in arXivdocs-target for

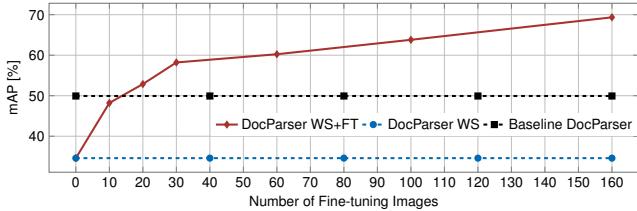


Figure 3: Performance of entity detection (mAP for IoU = 0.5) during fine-tuning.

| | IoU=0.5 | | | IoU=0.65 | | |
|--------------------|----------|-------|--------------|----------|-------|--------------|
| | Baseline | WS | WS+FT | Baseline | WS | WS+FT |
| All | 0.416 | 0.343 | 0.504 | 0.322 | 0.318 | 0.445 |
| <i>followed_by</i> | 0.413 | 0.387 | 0.506 | 0.314 | 0.366 | 0.447 |
| <i>parent_of</i> | 0.421 | 0.235 | 0.500 | 0.339 | 0.198 | 0.443 |
| Refined: | | | | | | |
| All | 0.453 | 0.382 | 0.615 | 0.363 | 0.354 | 0.558 |
| <i>followed_by</i> | 0.455 | 0.410 | 0.581 | 0.351 | 0.394 | 0.524 |
| <i>parent_of</i> | 0.451 | 0.317 | 0.679 | 0.389 | 0.263 | 0.620 |

Table 2: Performance in predicting hierarchical relations (as measured by F1).

the affected categories, coupled with an inferior model initialization, compared to DocParser WS+FT.

DocParser WS+FT outperforms the DocParser Baseline system across all measured IoU thresholds by a considerable margin. Using IoU thresholds above 0.5 leads to a performance decrease. Even though higher IoUs should generally correspond to better matches with the ground truth, they can penalize ambiguous cases and thus a correct detection. In sum, this confirms the effectiveness of our weak supervision in bolstering the overall performance.

Table 1 breaks down the performance by entity category. For DocParser WS+FT, we observe an especially good performance for detecting tabulars and figures. This is owed to the strong initialization of our system due to the high quality and large number of samples in our scalable weak supervision.¹¹

Figure 3 shows the fine-tuning. Only 20 fine-tuning samples are sufficient for DocParser WS+FT to surpass the baseline system DocParser (which is trained on 160 samples from the target dataset). It thus helps in reducing the labeling effort by a factor of around 8. Furthermore, we observe a steady increase in the performance of the fine-tuned networks with more samples. Notably, the highest performance increase is already achieved by the first 10 document images for fine-tuning.

¹¹For a few entities, the best performance is achieved a combination of the WS system together with a high IoU (e.g., BIBLIOGRAPHY BLOCK). A likely reason for this is the composition of arXivdocs-target. As BIBLIOGRAPHY entities were not specifically used as a criterion for the per-page sampling, fewer documents in the target dataset contained relevant entities, leading to decreased performance of the baseline and WS+FT systems.

| System | Schreiber et al. (2018) | Baseline | WS | WS+FT |
|--------|-------------------------|----------|--------|---------------|
| F1* | 0.9144 | 0.8443 | 0.8117 | 0.9292 |
| F1 | — | 0.8209 | 0.8056 | 0.9292 |

Table 3: ICDAR 2013 result on table structure parsing.

Notes: Evaluation of image-based systems on “ICDAR 50 %”, which uses a random subset containing 50 % of the competition set for testing. Schreiber et al. (2018) use a different, non-public 50 % random subset. Furthermore, Schreiber et al. (2018) choose the best system based on the test set as indicated by F1*. In contrast, F1 refers to the performance when the selection is based on the validation set.

Prediction of Hierarchical Relations Table 2 compares the classification of relations with and without post-processing. The best performance (across all Ψ) is achieved by DocParser WS+FT with an IoU of 0.5: it registers an F1 score of 0.615. Here, the use of weak supervision with fine-tuning yields consistent improvements. This is also due to the significant improvements of the prior entity detection for this system variant. In particular, for an IoU of 0.5, it outperforms the F1 score of the baseline system (F1 of 0.453) by 0.162. This amounts to a relative improvement of 35.8 %. Evidently, a smaller IoU threshold of 0.5 is beneficial. Higher IoU thresholds reduce the overall parsing performance as structure parsing builds on the prior detection of document entities.

The performance on hierarchical relations (F1 score of 0.615) is largely explained by our choice of a strict evaluation (i.e. the complete tuple including both entities must be correct). Overall, this performance is already highly effective in recovering the overall document structure. This is later confirmed as part of a qualitative assessment.

5.2 Robustness Check: Table Structure Parsing

Results: Table 3 compares the state-of-the-art for table structure parsing with our weak supervision strategy. Altogether, our weak supervision outperforms the state-of-the-art (Schreiber et al. 2018) by a considerable margin.

Discussion: Our system shows significant improvement over the image-based state of the art. We also compare our approach to the state-of-the-art heuristic-based system that operates on raw PDF files, instead of images, as input (Nurminen 2013). Even though our system does not utilize the additional information provided by raw PDF files, DocParser achieves an F1 score of 0.9292, compared to 0.9221 for the PDF-based system. We refrain from directly comparing the aforementioned F1 score with that from earlier experiments as the underlying target domains differ.

6 Related Work

OCR: Extracting text from document images has been extensively studied as part of optical character recognition (OCR) within the NLP community (e.g., Schäfer et al. 2011; Schäfer and Weitz 2012). To this end, the work by Katti et al. (2018) argued that OCR should be seen as a pre-processing step for downstream NLP tasks. As such, the au-

thors extract text-based information but not the hierarchical document structure as in our research.

Table Detection: Document renderings are commonly used for the task of table detection (rather than table structure parsing). Here, the objective is to predict the bounding boxes of tables, i.e., whether a pixel refers to a table or not (e.g., Yildiz, Kaiser, and Miksch 2005; Wang, Phillips, and Haralick 2004). Prior research on table detection has utilized data augmentation (Gilani et al. 2017), weak supervision (Li et al. 2019), and transfer learning (e.g., Siddiqui et al. 2018) to address the lack of large-scale domain-specific datasets. Similar to our research, efficient learning presents an issue for table detection. However, parsing of full pages requires effective identification of a much larger number of entities of multiple categories and high variety in shape per input.

Table Structure Parsing: There are works that recognize table structures from text or other syntactic tokens (Kieninger and Dengel 1998; Pivk et al. 2007) rather than directly from document renderings. As such, these works are tailored to tokens as input, and it is thus unclear how such an approach could theoretically be adapted to document renderings since our task inherently relies upon images as input. Because of the different input and thus the different datasets for benchmarking, the performance of the aforementioned works is not comparable to our approach. The works by Schreiber et al. (2018); Qasim, Mahmood, and Shafait (2019) draw upon deep neural networks to identify table structures for rendered inputs. However, they aim at a different purpose: parsing table structures, but not complete document hierarchies. As such, the authors do not attempt to identify text elements, nested figures, etc.

Weak Supervision for Document Layout: (Zhong, Tang, and Yepes 2019) use weak supervision for detection of page layout entities. The WS mechanism relies on matching external XML annotations with text extractions by a heuristic-based third-party tool. In contrast, our weak supervision directly builds on the L^AT_EXcompilation and can be readily extended to any new dataset of L^AT_EXsource files. Furthermore, the dataset features only 5 coarse categories and the system does not feature a relation classification component, thus being insufficient to acquire full document structures.¹²

Weak Supervision in NLP: Annotations in NLP are oftentimes costly and, as a result, there has been a recent surge in weak supervision. Weak supervision has now been applied to various tasks, such as text classification (e.g., Hingmire and Chakraborti 2014; Lin, He, and and Everson Richard 2011), information extraction (e.g., Hoffmann et al. 2011), and semantic parsing (e.g., Goldman et al. 2018). The methodological levers for obtaining weak labels are versatile and include, e.g., manual rules (e.g., Rabinovich et al. 2018), estimated models (e.g., Hoffmann et al. 2011), or reinforcement learning (Pröllochs, Feuerriegel, and Neumann 2019); however, not for document structure parsing.

7 Discussion and Conclusion

Efficiency: Our system requires only \sim 340 ms/document during entity detection (averaged over our validation set of

79 documents for DocParser WS+FT) on a single Titan Xp GPU with 12 GB VRAM and a batch size of 1. The relation detection in stage 2 only adds a minimal overhead of an average of 5.67 ms/document (10.81 ms/document with refinement) on a single CPU @ 2.1 GHz.

Qualitative Assessment: We performed a qualitative analysis on a subset of documents. We observe that, even for F1 scores below 0.5, the final document structure is often still very accurate. In fact, state-of-the-art OCR systems as natural baselines are outperformed significantly. This can be explained by our experiment design: we used very strict evaluation metrics. Hence, even small mismatches or ambiguities between the ground truth and predicted entities result in fairly large F1 penalties, despite high overall similarity. Details are in the supplements (including qualitative examples).

Detection Model Choice: Deep CNN models, including recent work (Tan and Le 2019; Duan et al. 2019), are heavily reliant on large training datasets. As such, we expect the impact of our technical contribution, as shown in our comparison of baseline and WS+FT models, to be the same across different modern CNN backbones. Our choice of Mask R-CNN as a tool for instance segmentation was also done in consideration of possible future extensions of DocParser to non-rectified documents. Here, the additional instance masks could guide the OCR or rectification process.

Future Work: In future work, we plan to explore approaches that can jointly learn entity and relation detection. Furthermore, we aim to further improve our system by enriching 2D inputs with textual features, e.g. high-dimensional word embeddings. The robustness of WS pre-training w.r.t. smaller subsets of arXivdocs-weak is another area of future investigation.

Conclusion: Despite the extensive interest of the NLP community in leveraging document structures (e.g., Apostolova and Tomuro 2014; Schäfer et al. 2011; Schäfer and Weitz 2012; Schreiber et al. 2018; Katti et al. 2018), the task of parsing complete document structures from renderings has been overlooked. To the best of our knowledge, we present the first system for this task. In particular, DocParser provides an effective alternative to state-of-the-art OCR which is still widespread in practice. In addition, DocParser allows to provide additional semantic input to downstream NLP tasks (e.g. information extraction).

8 Acknowledgments

Ce Zhang and the DS3Lab gratefully acknowledge the support from the Swiss National Science Foundation (Project Number 200021_184628), Innosuisse/SNF BRIDGE Discovery (Project Number 40B2-0_187132), European Union Horizon 2020 Research and Innovation Programme (DAPHNE, 957407), Botnar Research Centre for Child Health, Swiss Data Science Center, Alibaba, Cisco, eBay, Google Focused Research Awards, Oracle Labs, Swisscom, Zurich Insurance, Chinese Scholarship Council, and the Department of Computer Science at ETH Zurich.

¹²Additional comparison is included in the supplements.

References

- Abdulla, W. 2017. Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow.
- Antonacopoulos, A.; Bridson, D.; Papadopoulos, C.; and Pletschacher, S. 2009. A Realistic Dataset for Performance Evaluation of Document Layout Analysis. In *International Conference on Document Analysis and Recognition (ICDAR)*. ISBN 9780769537252. ISSN 15205363.
- Apostolova, E.; and Tomuro, N. 2014. Combining visual and textual features for information extraction from online flyers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1924–1929.
- Arif, S.; and Shafait, F. 2018. Table Detection in Document Images using Foreground and Background Features. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. ISBN 978-1-5386-6602-9.
- Chen, H.-H.; Tsai, S.-C.; and Tsai, J.-H. 2000. Mining Tables from Large Scale HTML Texts. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, 166–172. USA: Association for Computational Linguistics. ISBN 155860717X.
- Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 6569–6578.
- Embley, D. W.; Hurst, M.; Lopresti, D.; and Nagy, G. 2006. Table-processing Paradigms: A Research Survey.
- Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88(2): 303–338.
- Garncarek, Ł.; Powalski, R.; Stanisławek, T.; Topolski, B.; Halama, P.; and Graliński, F. 2020. LAMBERT: Layout-Aware language Modeling using BERT for information extraction. *arXiv preprint arXiv:2002.08087*.
- Gilani, A.; Qasim, S. R.; Malik, I.; and Shafait, F. 2017. Table Detection using Deep Learning. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*.
- Gobel, M.; Hassan, T.; Oro, E.; and Orsi, G. 2013. ICDAR 2013 Table Competition. In *International Conference on Document Analysis and Recognition (ICDAR)*. ISBN 978-0-7695-4999-6. ISSN 15205363.
- Goldman, O.; Latcinnik, V.; Nave, E.; Globerson, A.; and Berant, J. 2018. Weakly Supervised Semantic Parsing with Abstract Examples. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Govindaraju, V.; Zhang, C.; and Ré, C. 2013. Understanding Tables in Context Using Standard NLP Toolkits. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 658–664. Sofia, Bulgaria: Association for Computational Linguistics.
- He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*. ISBN 978-1-5386-0457-1. ISSN 0006-291X.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hingmire, S.; and Chakraborti, S. 2014. Sprinkling Topics For Weakly Supervised Text Classification. In *Annual Meeting of the ACL*.
- Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; and Weld, D. S. 2011. Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations. In *Annual Meeting of the ACL*.
- Hurst, M.; and Nasukawa, T. 2000. Layout and Language: Integrating Spatial and Linguistic Knowledge for Layout Understanding Tasks. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.
- Katti, A. R.; Reisswig, C.; Guder, C.; Brarda, S.; Bickel, S.; Höhne, J.; and Faddoul, J. B. 2018. Chagrid: Towards Understanding 2D Documents. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Kieninger, T.; and Dengel, A. 1998. The T-Recs Table Recognition and Analysis System. In *International Workshop on Document Analysis Systems (DAS)*.
- Laurens, J. 2008. Direct and reverse synchronization with SyncTEX. *TUGBoat* 29: 365–371.
- Li, M.; Cui, L.; Huang, S.; Wei, F.; Zhou, M.; and Li, Z. 2019. TableBank: Table Benchmark for Image-based Table Detection and Recognition. *arXiv preprint arXiv:1903.01949*.
- Lin, C.; He, Y.; and Everson Richard. 2011. Sentence Subjectivity Detection With Weakly-Supervised Learning. In *International Joint Conference on Natural Language Processing (IJCNLP)*.
- Lin, T. Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISBN 9781538604571.
- Lin, T. Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*. ISBN 978-3-319-10601-4. ISSN 16113349.
- Liu, X.; Gao, F.; Zhang, Q.; and Zhao, H. 2019. Graph Convolution for Multimodal Information Extraction from Visually Rich Documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, 32–39.
- Luong, M.-T.; Nguyen, T. D.; and Kan, M.-Y. 2012. Logical Structure Recovery in Scholarly Articles with Rich Document Features. In *Multimedia Storage and Retrieval Innovations for Digital Library Systems*, 270–292. IGI Global.

- Nurminen, A. 2013. *Algorithmic Extraction of Data in Tables in PDF Documents*. Master’s thesis, Tampere University of Technology.
- Pivk, A.; Cimiano, P.; Sure, Y.; Gams, M.; Rajković, V.; and Studer, R. 2007. Transforming Arbitrary Tables into Logical Form with TARTAR. *Data and Knowledge Engineering* 567–595. ISSN 0169023X.
- Pröllochs, N.; Feuerriegel, S.; and Neumann, D. 2019. Learning Interpretable Negation Rules via Weak Supervision at Document Level: A Reinforcement Learning Approach. In *NAACL-HLT*.
- Qasim, S. R.; Mahmood, H.; and Shafait, F. 2019. Rethinking table recognition using graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 142–147. IEEE.
- Rabinovich, E.; Sznajder, B.; Spector, A.; Shnayderman, I.; Aharonov, R.; Konopnicki, D.; and Slonim, N. 2018. Learning Concept Abstractness using Weak Supervision. In *EMNLP*.
- Rice, S. V.; Jenkins, F. R.; and Nartker, T. A. 1995. The Fourth Annual Test of OCR Accuracy. Technical report, Technical Report 95.
- Schäfer, U.; Kiefer, B.; Spurk, C.; Steffen, J.; and Wang, R. 2011. The ACL Anthology Searchbench. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations (ACL-HLT)*. Association for Computational Linguistics.
- Schäfer, U.; and Weitz, B. 2012. Combining OCR Outputs for Logical Document Structure Markup: Technical Background to the ACL 2012 Contributed Task. In *ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL ’12.
- Schreiber, S.; Agne, S.; Wolf, I.; Dengel, A.; and Ahmed, S. 2018. DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. In *International Conference on Document Analysis and Recognition (ICDAR)*. ISBN 9781538635865. ISSN 15205363.
- Siddiqui, S. A.; Malik, M. I.; Agne, S.; Dengel, A.; and Ahmed, S. 2018. DeCNT: Deep Deformable CNN for Table Detection. *IEEE Access* 74151–74161. ISSN 21693536.
- Tan, M.; and Le, Q. V. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.
- Tengli, A.; Yang, Y.; and Ma, N. L. 2004. Learning Table Extraction from Examples. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 987–993. Geneva, Switzerland: COLING.
- Wang, Y.; Phillips, I. T.; and Haralick, R. M. 2004. Table Structure Understanding and its Performance Evaluation. *Pattern Recognition* 1479–1497. ISSN 00313203.
- Yildiz, B.; Kaiser, K.; and Miksch, S. 2005. pdf2table: A Method to Extract Table Information from PDF Files. *2nd Indian International Conference on Artificial Intelligence (IICAI)*.
- Zanibbi, R.; Blostein, D.; and Cordy, J. 2004. A Survey of Table Recognition. *Document Analysis and Recognition* 1–33. ISSN 1433-2833.
- Zhong, X.; Tang, J.; and Yepes, A. J. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1015–1022. IEEE.

Appendix A Performance of Document Structure Parsing

A.1 Qualitative Evaluation

Figure 4 shows examples of parsed page structures that are generated by DocParser WS+FT.

We illustrate the effects of our structure-based refinement in Figure 5 and Figure 6. We observe that bounding boxes of parent entities from the raw predictions are refined such that they fully enclose all of their classified child entities. We particularly achieve improvement of the resulting predicted structure. For instance, for multi-figures, our refinement encloses figure graphics into individual figure structures to match the defined document grammar (see Figure 5). Figure 6 shows how two nested entities of the HEADING category are merged into a single entity during refinement.

We furthermore investigate the how the F1 measure for relation classification relates to overall parsing quality. Figure 7 depicts the detected entities and relations for a document with an F1 score of 0.267. We note that the overall quality of the parsed page is still high. Our relation classification requires entities in the page graph to be exactly matched with the corresponding entities in the ground truth by surpassing the IoU threshold. For instance, the detected HEADER entities are not matched with the ground truth, due to the shape mismatch. This causes a penalty to the F1 score, as several relation triples in the prediction that involve the headings are considered mismatches. Figure 8 shows another prediction with a low F1 score of 0.417. Here, mismatches can be accounted to the interpretation of entities that could be considered ambiguous. For instance, DocParser detects an inline heading in the last content block, while this text segment is interpreted as standard text in the ground truth. We additionally compare our results qualitatively to a state-of-the-art OCR software.¹³ We observe that the page region detection fails to differentiate between many of the considered semantic categories, e. g. HEADING HEADER and KEYWORDS in Figure 7 or EQUATION in Figure 8. We note that the OCR software has access to the original PDF files of full resolution and all meta information, while DocParser only operates on document renderings.

A.2 Reproducibility

For reproducibility purposes, we report results of DocParser on the validation set. Table 4 and Table 6 show the performance of the variants of DocParser for entity detection and prediction of hierarchical relations, respectively. Additionally, we include the complete results (including for IoU=0.8) on the test set in Table 5 and Table 7.

¹³We compare to outputs of ABBYY Finereader 15.

| AP | IoU=0.5 | | | IoU=0.65 | | | IoU=0.8 | | | |
|--------------|-------------------|-------------|--------------|-------------------|-------------|--------------|-------------------|-------------|-------------|-------------|
| | Baseline WS WS+FT | | | Baseline WS WS+FT | | | Baseline WS WS+FT | | | |
| | mAP | 50.1 | 41.2 | 71.0 | 37.2 | 37.9 | 59.1 | 15.6 | 24.9 | 37.7 |
| abstract | 78.1 | 99.6 | 100.0 | 78.1 | 81.9 | 100.0 | 51.0 | 52.1 | 66.3 | |
| affiliation | 50.7 | 0.0 | 57.6 | 39.1 | 0.0 | 39.2 | 1.5 | 0.0 | 1.9 | |
| author | 26.2 | 0.0 | 31.1 | 24.0 | 0.0 | 18.8 | 0.0 | 0.0 | 5.5 | |
| bib. block | 27.6 | 75.1 | 78.8 | 14.3 | 75.1 | 81.8 | 14.3 | 57.1 | 60.3 | |
| cont. block | 82.9 | 69.9 | 90.7 | 76.1 | 65.2 | 86.9 | 57.8 | 54.0 | 73.6 | |
| date | 0.0 | 0.0 | 34.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| equation | 59.8 | 54.2 | 83.4 | 35.9 | 46.7 | 73.9 | 11.6 | 30.8 | 39.8 | |
| fig. caption | 50.7 | 39.8 | 70.0 | 50.6 | 14.5 | 62.3 | 22.4 | 18.1 | 44.6 | |
| fig. graphic | 40.3 | 5.4 | 77.0 | 31.7 | 7.5 | 74.5 | 12.0 | 4.5 | 52.8 | |
| figure | 71.1 | 39.3 | 53.9 | 60.0 | 40.8 | 51.6 | 18.5 | 17.1 | 41.7 | |
| footer | 62.1 | 0.0 | 74.5 | 35.2 | 0.0 | 56.6 | 26.2 | 0.0 | 28.8 | |
| header | 65.0 | 0.0 | 68.1 | 45.9 | 0.0 | 44.5 | 9.4 | 0.0 | 12.2 | |
| heading | 56.8 | 58.9 | 72.6 | 38.3 | 54.8 | 65.8 | 7.8 | 21.1 | 21.4 | |
| item | 0.0 | 68.6 | 69.6 | 0.0 | 68.7 | 63.7 | 0.0 | 36.4 | 33.9 | |
| itemize | 0.0 | 69.1 | 63.5 | 0.0 | 63.5 | 53.2 | 0.0 | 33.3 | 44.7 | |
| keywords | 40.7 | 0.0 | 48.5 | 18.5 | 0.0 | 15.1 | 0.0 | 0.0 | 12.4 | |
| page nr. | 57.5 | 0.0 | 60.3 | 18.7 | 0.0 | 23.2 | 0.1 | 0.0 | 2.4 | |
| tab. caption | 56.4 | 81.2 | 91.5 | 38.1 | 88.7 | 89.1 | 9.6 | 50.8 | 44.2 | |
| table | 92.0 | 94.6 | 97.4 | 62.9 | 88.3 | 87.6 | 18.7 | 69.3 | 77.5 | |
| tabular | 83.0 | 67.7 | 96.3 | 76.2 | 62.1 | 93.2 | 50.8 | 52.6 | 89.3 | |

Table 4: Validation set: Comparison of entity detection (average precision) without structure-based refinement.

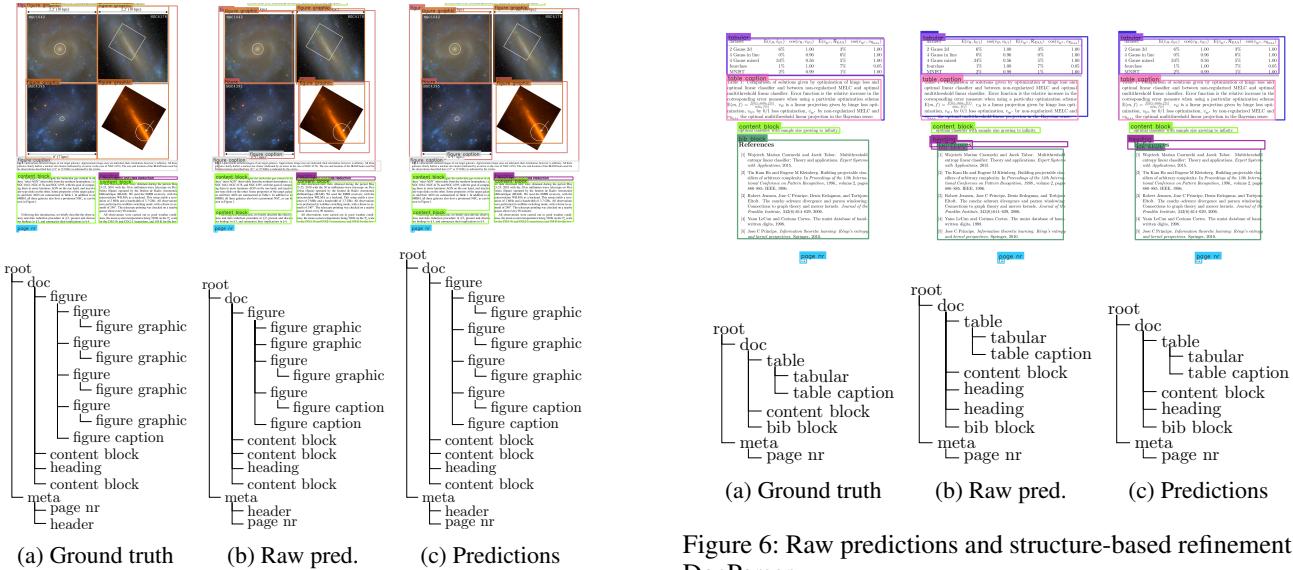


Figure 5: Raw predictions and structure-based refinement in DocParser.

Appendix D Datasets with Document Structure: arXivdocs-weak

Figure 10 and Table 10 show the descriptive statistics of the dataset. Evidently, the most common category in the dataset is content line. Content lines typically represent leaf nodes in the graph and are children of larger entities, such as abstract, captions, or content blocks.

Component 4: Structure-Based Refinement

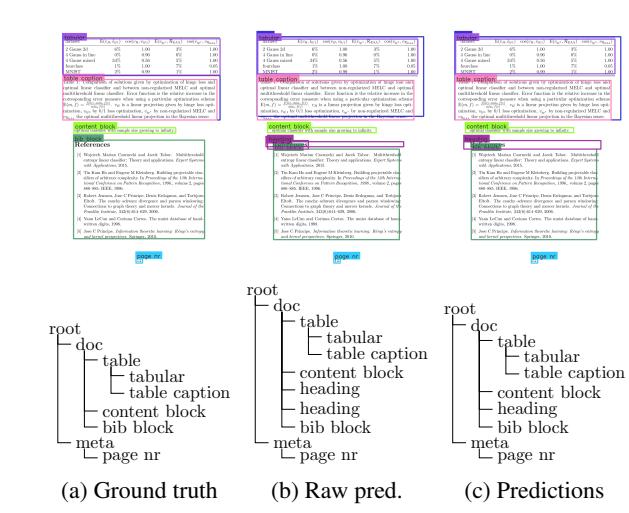


Figure 6: Raw predictions and structure-based refinement in DocParser.

(4) In our experiments, we use $r = 30$ for structure-based refinement. During development, we observed only minor differences for values of $r = 10$ and higher. To confirm this, we analyze the performance of DocParser WS+FT for $r = 2, 5, 10, 20, 30$ on the validation set (see Table 11).¹⁴ Here we observe that the accuracy of our system remains unchanged for values of $r \geq 10$.

¹⁴Note that results are given for a single model and can differ from the detailed relation classification evaluation, where we average over three models.

| AP | IoU=0.5 | | | IoU=0.65 | | | IoU=0.8 | | |
|--------------|-------------------|-------------|--------------|-------------------|-------------|-------------|-------------------|-------------|-------------|
| | Baseline WS WS+FT | | | Baseline WS WS+FT | | | Baseline WS WS+FT | | |
| | mean AP | 49.9 | 34.6 | 69.4 | 38.5 | 32.4 | 56.5 | 14.7 | 23.5 |
| abstract | 95.2 | 90.5 | 95.2 | 90.5 | 81.0 | 95.2 | 48.9 | 28.3 | 75.1 |
| affiliation | 51.6 | 0.0 | 46.0 | 5.9 | 0.0 | 16.2 | 1.0 | 0.0 | 0.8 |
| author | 18.0 | 0.0 | 23.6 | 20.4 | 0.0 | 16.7 | 4.9 | 0.0 | 8.0 |
| bib. block | 42.4 | 79.1 | 94.7 | 43.2 | 93.9 | 80.3 | 12.7 | 96.2 | 65.6 |
| cont. block | 89.3 | 69.8 | 88.4 | 83.2 | 67.0 | 84.4 | 64.4 | 55.6 | 74.2 |
| date | 0.0 | 0.0 | 24.1 | 0.0 | 0.0 | 9.3 | 0.0 | 0.0 | 0.0 |
| equation | 65.8 | 54.5 | 82.1 | 40.6 | 52.1 | 72.8 | 8.9 | 36.4 | 38.4 |
| fig. caption | 47.8 | 30.5 | 69.2 | 44.0 | 17.7 | 59.5 | 16.7 | 19.6 | 39.8 |
| fig. graphic | 22.3 | 5.2 | 60.2 | 15.9 | 4.4 | 54.5 | 6.2 | 1.6 | 36.6 |
| figure | 47.8 | 35.3 | 63.5 | 44.0 | 33.9 | 59.4 | 22.5 | 21.0 | 51.3 |
| footer | 55.7 | 0.0 | 69.3 | 48.9 | 0.0 | 59.7 | 5.0 | 0.0 | 7.9 |
| header | 79.7 | 0.0 | 88.3 | 64.8 | 0.0 | 56.6 | 12.1 | 0.0 | 6.5 |
| heading | 53.7 | 52.1 | 66.4 | 33.1 | 46.0 | 45.4 | 6.7 | 26.3 | 16.7 |
| item | 0.0 | 33.6 | 50.5 | 0.0 | 35.3 | 33.5 | 0.0 | 53.0 | 10.4 |
| itemize | 0.0 | 25.0 | 58.3 | 0.0 | 25.0 | 50.0 | 0.0 | 0.0 | 58.3 |
| keywords | 36.4 | 0.0 | 59.0 | 36.4 | 0.0 | 43.0 | 20.5 | 0.0 | 22.3 |
| page nr. | 74.7 | 0.0 | 77.3 | 28.5 | 0.0 | 42.0 | 0.8 | 0.0 | 2.0 |
| tab. caption | 55.2 | 69.1 | 76.6 | 40.2 | 61.6 | 63.4 | 16.5 | 41.5 | 28.6 |
| table | 84.5 | 96.3 | 94.3 | 62.7 | 87.9 | 89.6 | 14.6 | 68.2 | 80.2 |
| tabular | 78.4 | 50.8 | 100.0 | 68.4 | 42.4 | 99.5 | 32.4 | 22.3 | 89.0 |

Table 5: Test set: Average precision (AP) of entity detection.

| | IoU=0.5 | | | IoU=0.65 | | | IoU=0.8 | | |
|--------------------|-------------------|-------|--------------|-------------------|-------|--------------|-------------------|-------|--------------|
| | Baseline WS WS+FT | | | Baseline WS WS+FT | | | Baseline WS WS+FT | | |
| | All | 0.406 | 0.369 | 0.550 | 0.337 | 0.350 | 0.478 | 0.131 | 0.221 |
| <i>followed_by</i> | 0.360 | 0.368 | 0.531 | 0.303 | 0.356 | 0.454 | 0.119 | 0.216 | 0.294 |
| <i>parent_of</i> | 0.519 | 0.372 | 0.592 | 0.418 | 0.333 | 0.532 | 0.160 | 0.235 | 0.379 |

| Refined: | IoU=0.5 | | | IoU=0.65 | | | IoU=0.8 | | |
|--------------------|-------------------|-------|--------------|-------------------|-------|--------------|-------------------|-------|--------------|
| | Baseline WS WS+FT | | | Baseline WS WS+FT | | | Baseline WS WS+FT | | |
| | All | 0.447 | 0.401 | 0.658 | 0.359 | 0.385 | 0.576 | 0.164 | 0.266 |
| <i>followed_by</i> | 0.402 | 0.390 | 0.602 | 0.310 | 0.384 | 0.513 | 0.140 | 0.254 | 0.338 |
| <i>parent_of</i> | 0.551 | 0.428 | 0.771 | 0.470 | 0.387 | 0.704 | 0.217 | 0.298 | 0.555 |

Table 6: Validation set: Performance in predicting hierarchical relations (as measured by F1).

We additionally provide pseudo-code for our refinement procedure in Algorithm 1.

Component 5: Scalable Weak Supervision

To analyze the degree of noise in arXivdocs-weak, we evaluate the average precision for the weak annotations against the manually generated ground truth in arXivdocs-target. Table 12 shows the accuracies of arXivdocs-weak for different IoU values, as measured on the training split of arXivdocs-target. We observe various AP values of 0, indicating the absence of the respective categories in arXivdocs-weak. Furthermore, for the majority of categories, the measured is relatively low (AP ≤ 0.5 for IoU ≥ 0.5). This emphasizes the systematic noise in arXivdocs-weak and confirms the positioning of our experimental setting in the domain of weak supervision.

Appendix E Computational Setup

Mask R-CNN Our used Mask R-CNN model is illustrated in Figure 11.

Appendix F Related Work

F.1 Weak Supervision for Document Layout:

(Zhong, Tang, and Yipes 2019) (PN) use weak supervision for detection of page layout entities. The dataset features 5 coarse categories, compared to 23 fine-grained categories in arXivdocs. Furthermore the system does not contain a relation classification component. Following, we examine differences and correspondences between the five classes in (Zhong, Tang, and Yipes 2019) and our arXivdocs:

- TEXT: Corresponds to CONTENT BLOCK in arXivdocs. In contrast to our dataset, the TEXT category corresponds to individual paragraphs (instead of uninterrupted text on a single column) and is used for captions.
- TITLE: corresponds to our HEADER category.
- LIST: corresponds to our ITEMIZE category. A difference here is that LIST entities in PN are separated by columns.
- TABLE: corresponds to our TABULAR. In contrast to arXivdocs, they do not feature nesting relations that contain,

| | IoU=0.5 | | | IoU=0.65 | | | IoU=0.8 | | |
|--------------------|----------|-------|--------------|----------|-------|--------------|----------|-------|--------------|
| | Baseline | WS | WS+FT | Baseline | WS | WS+FT | Baseline | WS | WS+FT |
| All | 0.416 | 0.343 | 0.504 | 0.322 | 0.318 | 0.445 | 0.114 | 0.214 | 0.313 |
| <i>followed_by</i> | 0.413 | 0.387 | 0.506 | 0.314 | 0.366 | 0.447 | 0.118 | 0.244 | 0.308 |
| <i>parent_of</i> | 0.421 | 0.235 | 0.500 | 0.339 | 0.198 | 0.443 | 0.103 | 0.138 | 0.325 |
| Refined: | | | | | | | | | |
| All | 0.453 | 0.382 | 0.615 | 0.363 | 0.354 | 0.558 | 0.157 | 0.254 | 0.395 |
| <i>followed_by</i> | 0.455 | 0.410 | 0.581 | 0.351 | 0.394 | 0.524 | 0.150 | 0.277 | 0.352 |
| <i>parent_of</i> | 0.451 | 0.317 | 0.679 | 0.389 | 0.263 | 0.620 | 0.172 | 0.202 | 0.474 |

Table 7: Test set: Performance in predicting hierarchical relations (as measured by F1).

```

Data: Detected Entities
Result: Refined Entities; Hierarchical Relations
counter=0;
while counter  $\leq \theta$  do
    counter++;
    Classify all hierarchical relations;
    if (1) Parent entity bounding boxes don't fully
    enclose children then
        Expand parent bounding boxes s.t. they
        enclose children;
        Go to start of loop;
    end
    if (2) Directly nested entities of same category
    exist then
        Merge directly nested entities into a single
        entity;
        Go to start of loop;
    end
    if (3) Siblings found that are not allowed to
    co-exist in hierarchy then
        Enclose groups of siblings with new, valid
        parent entities;
        Go to start of loop;
    end
    if (4) Possible parent found in neighborhood of a
    parent-less entity then
        Expand matched parents bounding boxes to
        enclose child entities;
        Go to start of loop;
    end
    Exit loop;
end
Classify all hierarchical relations;
Algorithm 1: Structure-based refinement.
```

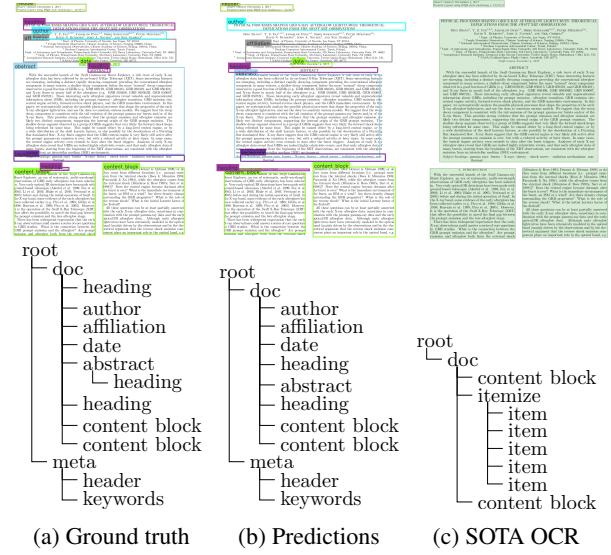


Figure 7: Output with low F1 score (0.267), compared to state-of-the-art OCR software. Green regions in the OCR page recognition correspond to “text areas”. Using the OCR tool, we convert the page to HTML and use our tree-graph to represent the resulting structure. The affiliation section is converted into a list by the OCR software during conversion to HTML.

for instance, TABLE CAPTION entities. Fine-grained children, such as cells, rows and columns are also not featured.

- FIGURE: corresponds roughly to the concept of FIGURE GRAPHIC in arXivdocs. However, no nesting relations (i.e. sub-figures) or captions are featured.

We evaluate the feasibility of using the dataset presented in PN for pre-training. We use the same pre-training procedure as in our experiments that utilize arXivdocs. To account for the difference of pre-training and target domains, we use an extended fine-tuning procedure of PN that matches the pre-training scheme of up to 80,000 iterations. Table 13 shows results for entity detection. Here we observe that pre-training improves the performance of the system, when compared to DocParser Baseline that does not use weak supervision. We also observe that pre-training with the PN dataset results in

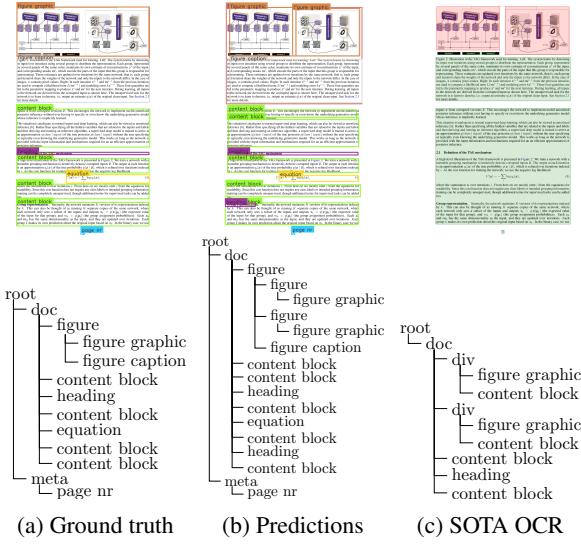


Figure 8: Output with low F1 score (0.417), compared to state-of-the-art OCR software. Green and red regions in the OCR page recognition correspond to “text areas” and “picture areas”, respectively. Using the OCR tool, we convert the page to HTML and use our tree-graph to represent the resulting structure. The content block in the second “div” section corresponds to the full figure caption text.

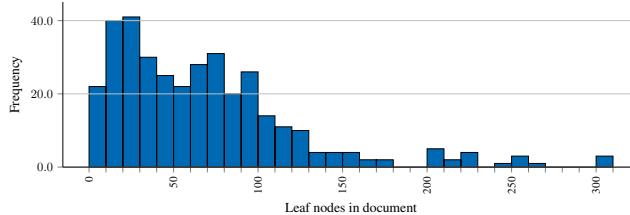


Figure 9: Number of leaf nodes in documents.

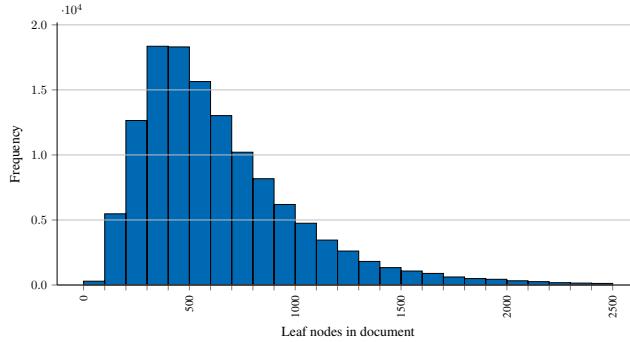


Figure 10: Number of leaf nodes in documents for weak supervision.

significantly lower mAP values, e.g. 60.0 after fine-tuning compared to 69.4 in DocParser WS+FT at IoU = 0.5. For some entity categories, we observe higher individual AP values for PN, e.g. AFFILIATION at IoU = 0.5. We attribute this to the higher occurrence of more compact TEXT entities in

| Entity (\mathcal{C}) | Relation types Ψ | Valid entities | Notes |
|--------------------------|-----------------------|--------------------|-----------|
| Abstract | <i>parent_of</i> | Heading | |
| Figure | <i>parent_of</i> | Figure | Float |
| | <i>parent_of</i> | Fig. graphic | |
| | <i>parent_of</i> | Fig. caption | |
| Fig. graphic | <i>followed_by</i> | Fig. caption | if nested |
| Item | <i>parent_of</i> | Equation | |
| Itemize | <i>parent_of</i> | Item | |
| Table | <i>parent_of</i> | Tabular | Float |
| | <i>parent_of</i> | Tab. caption | |
| Tabular | <i>parent_of</i> | Tab. cell | |
| | <i>parent_of</i> | Tab. row | |
| | <i>parent_of</i> | Tab. col. | |
| Date | null | — | Meta |
| Footer | null | — | Meta |
| Header | null | — | Meta |
| Keywords | null | — | Meta |
| PageNr | null | — | Meta |
| All others | <i>parent_of</i> , | — | |
| Any entity | <i>followed_by</i> | <i>any sibling</i> | |

Table 8: Document grammar for different entity categories that is utilized in our heuristics. Every category can by default exist on the highest hierarchical level, i.e., without being nested. Hierarchical nesting for the child entities of floats, e.g. captions, is, however, encouraged in the automatic refinement process. Further details are included in the supplements.

PN. Additionally, this could also be caused by our experimental protocol in which early stopping is applied function of the mAP value, instead of individual AP values. As such, there is a performance trade-off between individual entity categories.

Appendix G Robustness Check: Table Structure Parsing

We perform robustness checks of DocParser on the table structure parsing task. DocParser is evaluated for entity detection on arXivdocs-target and structure parsing on the ICDAR 2013 table structure dataset.

We received the outputs for the ICDAR “competition” dataset from the authors of (Nurminen 2013). We used the evaluation script provided by the competition organizers to calculate the ICDAR 50 % performance.

We match our table cell predictions with the text element locations provided by (Nurminen 2013) in order to generate XML files that are compared to the ground truth by the scripts provided on the competition website. Matches are determined by the fraction of overlap between cell and text bounding boxes $\gamma = \frac{\text{area}(B_{\text{cell}} \cap B_{\text{text}})}{\text{area}(B_{\text{text}})}$, using $\gamma \geq 0.5$.

G.1 Table Structure Heuristics

For the ordering of table structure entities, we draw upon a set of special heuristics. The reason for this is that nesting relationships are often too complex to model with the previously described parent-child relationships, e.g. for cells

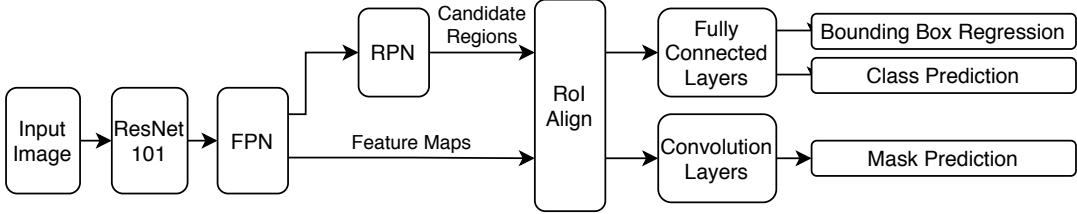


Figure 11: Mask R-CNN overview.

| Category | Frequency | % | Avg. depth |
|------------------|-----------|-------|------------|
| abstract | 63 | 0.20 | 1.00 |
| affiliation | 82 | 0.26 | 1.00 |
| author | 89 | 0.28 | 1.00 |
| bibliogr. block | 32 | 0.10 | 2.00 |
| bibliography | 24 | 0.08 | 1.08 |
| code | 3 | 0.01 | 2.00 |
| content block | 1009 | 3.23 | 2.05 |
| content line | 10,729 | 34.33 | 3.06 |
| content lines | 627 | 2.01 | 3.04 |
| date | 16 | 0.05 | 1.00 |
| equation | 353 | 1.13 | 1.97 |
| equation formula | 364 | 1.16 | 3.01 |
| equation label | 275 | 0.88 | 2.95 |
| figure | 607 | 1.94 | 2.44 |
| figure caption | 404 | 1.29 | 3.16 |
| figure graphic | 454 | 1.45 | 3.63 |
| footer | 81 | 0.26 | 1.00 |
| header | 106 | 0.34 | 1.01 |
| heading | 398 | 1.27 | 2.10 |
| item | 63 | 0.20 | 3.08 |
| itemize | 24 | 0.08 | 1.96 |
| page nr | 261 | 0.84 | 1.00 |
| section | 527 | 1.69 | 1.29 |
| keywords | 36 | 0.12 | 1.00 |
| table | 185 | 0.59 | 1.81 |
| table caption | 183 | 0.59 | 2.80 |
| table cell | 11,146 | 35.67 | 3.54 |
| table col | 1109 | 3.55 | 3.69 |
| table row | 1812 | 5.80 | 3.68 |
| tabular | 187 | 0.60 | 2.80 |

Table 9: Statistics by entity of arXivdocs-target.

belonging to multiple rows and/or columns. Due to these complex relations, bottom-up creation of table row and table column entity bounding boxes from associated children is also challenging. We, therefore, generate rows, columns, and cells on the same hierarchical levels and store structure information in an additional attribute in each entity.

The following heuristics are applied:

1. Rows are sorted, based on the y -coordinate of their centroids. Columns are analogously sorted, based on their centroid x -coordinates.
2. Row entities that are located such that their bounding box is fully contained inside the bounding box of other row entities are determined. All such direct nestings are resolved as follows: (1) If a row entity contains exactly one

| Category | Frequency | % | Avg. Depth |
|------------------|------------|-------|------------|
| abstract | 89,291 | 0.09 | 1.00 |
| author | 48 | 0.00 | 3.00 |
| bibliogr. block | 242,412 | 0.26 | 2.94 |
| bibliography | 80,864 | 0.09 | 1.93 |
| caption | 26 | 0.00 | 4.38 |
| content block | 5,033,714 | 5.32 | 2.41 |
| content line | 63,339,623 | 66.96 | 3.47 |
| date | 5 | 0.00 | 2.00 |
| equation | 1,489,078 | 1.57 | 2.44 |
| equation formula | 1,743,705 | 1.84 | 3.43 |
| equation label | 1,503,778 | 1.59 | 3.44 |
| figure | 478,086 | 0.51 | 2.50 |
| figure caption | 263,495 | 0.28 | 3.48 |
| figure graphic | 408,088 | 0.43 | 3.52 |
| heading | 975,414 | 1.03 | 2.54 |
| item | 436,222 | 0.46 | 3.58 |
| itemize | 140,415 | 0.15 | 2.64 |
| meta | 127,477 | 0.13 | 3.00 |
| section | 1,296,707 | 1.37 | 1.56 |
| table | 292,110 | 0.31 | 2.43 |
| table caption | 206,215 | 0.22 | 3.42 |
| table cell | 12,343,327 | 13.05 | 4.40 |
| table col | 1,285,945 | 1.36 | 4.42 |
| table row | 2,533,799 | 2.68 | 4.41 |
| tabular | 280,572 | 0.30 | 3.43 |
| title | 16 | 0.00 | 3.00 |

Table 10: Summary statistics by entity of arXivdocs-weak dataset.

other row entity, remove the contained entity. (2) Remove row entities that contain more than one other row entity. Analogously, we proceed to discard column entities with direct nesting.

3. The bounding box (i.e., “union”) of all row and column entities is computed. However, the size of this bounding box might differ from the bounding boxes of the row and column entities. Hence, the bounding boxes of all rows are adjusted so that all adjacent rows have the width as the “union”. Analogously, the height for all bounding boxes belonging to columns are adjusted.
4. The location of rows might not be located at the center of adjacent rows. This is achieved by setting the y -coordinate of each row to the average of its adjacent rows. An analogous adjustment is performed for the x -coordinates of columns.

| <i>r</i> | IoU=0.5 | IoU=0.65 | IoU=0.8 |
|----------|---------|----------|---------|
| 2 | 0.619 | 0.563 | 0.384 |
| 5 | 0.680 | 0.619 | 0.427 |
| 10 | 0.680 | 0.619 | 0.427 |
| 20 | 0.680 | 0.619 | 0.427 |
| 30 | 0.680 | 0.619 | 0.427 |

Table 11: Impact of r on the relation classification performance on the development set for a DocParser WS+FT model.

| AP | IoU=0.5 | IoU=0.65 | IoU=0.8 |
|--------------|---------|----------|---------|
| mAP | 30.0 | 25.5 | 19.4 |
| abstract | 61.1 | 43.5 | 9.8 |
| affiliation | 0.0 | 0.0 | 0.0 |
| author | 0.0 | 0.0 | 0.0 |
| bib. block | 47.6 | 47.6 | 31.6 |
| cont. block | 38.9 | 32.7 | 23.2 |
| date | 0.0 | 0.0 | 0.0 |
| equation | 26.2 | 24.5 | 22.8 |
| fig. caption | 24.9 | 23.8 | 23.8 |
| fig. graphic | 18.5 | 14.7 | 14.7 |
| figure | 26.5 | 19.5 | 11.2 |
| footer | 0.0 | 0.0 | 0.0 |
| header | 0.0 | 0.0 | 0.0 |
| heading | 33.6 | 33.6 | 33.6 |
| item | 25.1 | 9.6 | 4.8 |
| itemize | 54.5 | 42.4 | 42.4 |
| keywords | 0.0 | 0.0 | 0.0 |
| page nr. | 0.0 | 0.0 | 0.0 |
| tab. caption | 76.3 | 76.3 | 72.7 |
| table | 74.7 | 60.5 | 44.6 |
| tabular | 91.7 | 80.8 | 52.1 |

Table 12: Average precision (AP) of entities in arXivdocs-weak, compared to the training split of arXivdocs-target

- Row and column numbers are assigned to separately detected cells as follows: for all cell entities from DocParser, we calculate the overlap between the vertical cell border and all vertical row borders. We then calculate the rows for which the length of the overlap is equal or larger than 50 % of the height of a row. The number of the corresponding row is then assigned to the row range of the cell. Analogously, we match cells to columns based on their horizontal overlap. If a cell is matched with more than one row or column, its bounding box is adjusted such that its borders lie on the grid of row and column borders. All other cells without assignment are dismissed.
- A grid of rectangular cells is generated from the intersection of all rows and columns for all positions in the table where no multi-row or multi-column cell exists.

G.2 Implementation Details

Entity Detection We use the hierarchical document annotations in arXivdocs-weak to identify 222,195 table structure entities that are used for weak supervision. The corresponding cropped tabular regions and their child entities,

i.e., rows, columns, and cells, are used as training input for the specialized system. The sampling process is stratified to bolster prediction performance: we use all row and column annotations, but only a subset of all table cell annotations. The reason is that regular cells can be reconstructed from robust detections of rows and columns. Row and column detection performance can, however, be adversely affected by category imbalance during sampling. The comparably large number of individual table cells per input creates such imbalance. Therefore, we only sample table cells that appear in the first table row and column, as well as cells spanning multiple rows or columns. Altogether, this aids the detection of multi-row and -column cells. Again, these cells can not be robustly reconstructed from regular rows and columns otherwise. The parameters for entity samples per image, ground truth samples per image and maximum number of predictions per image are set to 200, 200 and 400, respectively.

The train, validation and test splits of arXivdocs-target contain 87, 39, and 61 tabular entities, respectively. Crops of the entities are used for training and evaluation of the system specialized for table structure.

ICDAR 2013 Table Structure Dataset: The ICDAR 2013 table structure dataset (Gobel et al. 2013) is designed to evaluate table structure parsing. This dataset is later leveraged as part of our robustness check so that we can evaluate our weak supervision against state-of-the-art approaches for structure parsing. The dataset consists of 123 images, for which structure annotations, including cells, rows, and columns were created. The dataset comes without predefined train/test split; hence, we follow Schreiber et al. (2018) and split the so-called “competition” part of the dataset with a 50%/50%-ratio. One of the splits is used for evaluation. The other split is used in addition to the so-called “practice” part of the dataset for training and validation. We follow the official competition rules from ICDAR 2013 as follows: we operate directly on table sub-regions and thus create individual cropped images of these regions for training, validation, and evaluation. We generate rectangular row and column bounding boxes from the provided cell bounding boxes and their respective row- and column ranges. The resulting rows and columns are then further modified as follows: A tabular bounding box is determined as union bounding box of all cells. Bounding boxes of rows that share a border with the outer tabular are extended such that their borders fully align with the tabular. Afterwards, we move the borders of all pairs of neighboring rows to their respective midpoint. Analogously, we adjust all column bounding boxes. Cell bounding boxes are newly created from row and column intersections in a final step.

Entity Detection on arXivdocs-target Analogously to our evaluation on full documents, we measure mAP for table rows and table columns on a subset of table regions in arXivdocs-target. Average precision for joint detection of gg table rows and columns and the impact of fine-tuning are shown in Figure 12. Compared to full document pages, we measure higher mAPs for all systems. We observe that the weakly supervised model outperforms DocParser Baseline without having been trained on the target domain. We ob-

| AP | IoU=0.5 | | IoU=0.65 | | IoU=0.8 | |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | WS(PN) | WS+FT(PN) | WS(PN) | WS+FT(PN) | WS(PN) | WS+FT(PN) |
| | mean AP | 8.4 | 60.0 | 3.8 | 48.3 | 2.5 |
| abstract | 0.0 | 89.2 | 0.0 | 80.5 | 0.0 | 50.2 |
| affiliation | 0.0 | 63.8 | 0.0 | 35.8 | 0.0 | 4.8 |
| author | 0.0 | 34.3 | 0.0 | 22.0 | 0.0 | 1.1 |
| bib. block | 0.0 | 60.3 | 0.0 | 60.3 | 0.0 | 45.5 |
| cont. block | 34.9 | 90.3 | 20.1 | 86.8 | 14.5 | 76.1 |
| date | 0.0 | 16.7 | 0.0 | 0.0 | 0.0 | 0.0 |
| equation | 0.0 | 77.3 | 0.0 | 57.1 | 0.0 | 18.0 |
| fig. caption | 0.0 | 64.8 | 0.0 | 62.4 | 0.0 | 27.9 |
| fig. graphic | 0.0 | 38.4 | 0.0 | 33.2 | 0.0 | 19.8 |
| figure | 23.0 | 49.9 | 6.0 | 46.3 | 1.5 | 38.0 |
| footer | 0.0 | 69.1 | 0.0 | 56.4 | 0.0 | 3.9 |
| header | 0.0 | 76.9 | 0.0 | 63.7 | 0.0 | 5.7 |
| heading | 27.4 | 63.8 | 12.3 | 46.4 | 2.9 | 22.7 |
| item | 0.0 | 1.7 | 0.0 | 1.7 | 0.0 | 0.0 |
| itemize | 36.5 | 25.0 | 12.5 | 0.0 | 25.0 | 0.0 |
| keywords | 0.0 | 50.0 | 0.0 | 48.9 | 0.0 | 42.4 |
| page nr. | 0.0 | 80.3 | 0.0 | 38.0 | 0.0 | 1.0 |
| tab. caption | 0.0 | 62.7 | 0.0 | 50.0 | 0.0 | 21.1 |
| table | 46.5 | 90.8 | 25.0 | 82.1 | 5.4 | 59.8 |
| tabular | 0.0 | 94.8 | 0.0 | 94.7 | 0.0 | 79.4 |

Table 13: Average precision (AP) of entity detection on the test set, using (Zhong, Tang, and Yepes 2019) (PN) and structure-based refinement.

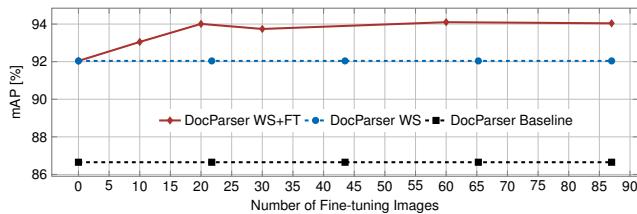


Figure 12: Comparison of test mAP (IoU=0.5) for three variants of DocParser for detection of table structure annotations. We follow the same procedure as described for fine-tuning with the default document entities. The weakly supervised system DocParser WS outperforms the baseline system without fine-tuning (FT). Fine-tuning with 10 or more images yields additional performance gains.

serve additional significant performance improvements in DocParser WS systems that were fine-tuned with 10 to 87 images. Because of the intricacies evaluating hierarchical structure parsing for tables, we perform a separate evaluation of DocParser for this task.