You've translated it, now what?

Michael Maxwell Shabnam Tafreshi Aquia Richburg Balaji Kodali Kymani Brown mmaxwell@arlis.umd.edu stafreshi@arlis.umd.edu arichbu1@umd.edu bkodali@umd.edu krown001@terpmail.umd.edu

Abstract

Humans use document formatting to discover important phrases and document structure. But when machines process a paper–especially documents OCRed from images–these cues are often invisible to downstream processes: words in footnotes or body text are treated as just as important as words in titles. It would be better for indexing and summarization tools to be guided by implicit document structure.

In an ODNI-sponsored project, the Applied Research Laboratory for Intelligence and Security (ARLIS) looked at inferring document structure from the formatting in OCRed text. Most OCR engines output results as hOCR (an XML format), giving bounding boxes around characters. In theory, this also provides style information such as bolding and italicization, but in practice, this capability is limited. For example, the Tesseract OCR tool provides bounding boxes, but does not attempt to detect bold or italicized text (relevant to author emphasis and specialized fields in e.g. print dictionaries).

Our project inferred font size from hOCR bounding boxes, and using that and other cues (e.g. the fact that titles tend to be short) determined which text constituted section titles; from this, a document outline can be created. We also experimented with algorithms for detecting bold text. Our best algorithm has a much improved recall and precision, although the exact numbers are font-dependent.

The next step is to incorporate inferred structure into the output of machine translation. One method is to embed XML tags for inferred structure into the text extracted from the imaged document, and to either pass the strings enclosed by XML tags to the MT engine individually, or pass the tags through the MT engine without modification. This structural information can guide downstream bulk processing tasks such as summarization and search, and also enables building tables of contents for human users examining individual translated documents.

1 Introduction

As you decided whether to read this paper, you probably read the **title** first, then maybe the **abstract**. You might have looked at the **pictures** to see if they whetted your interest, and looked at the **title of this section** before starting to read the section itself.

You may also have noticed the bolding in the paragraph above, and thought that perhaps those were clues to what the paper is about—which of course would be *right*.

In short, you will undoubtedly made use of **the paper's formatting**, both to decide whether reading a paper is worth your time, and if so, to better understand what the paper is about.

Now suppose you have a PDF of a paper written in a language you don't know, or an image of a document that has to be OCRed, or a paper document. You extract the text and pass it through a machine translation system, and out comes translated text—and if you're lucky, the translation is both fluent and accurate. But where is the formatting that you found so helpful?

2 The Problem

Documents are often written using overt or covert markup, where the markup more or less explicitly defines the document structure. Overt structural markup is used in formats like the following:

- DocBook (Norman and Hamilton 2010), which defines structural tags for technical documentation, for example <para> (paragraph), <qandaentry> (a question-and-answer, or Q-and-A, entry), and <guimenuitem> (the name of a terminal menu item in a GUI).
- Text Encoding Initiative (TEI, https://tei-c.org), defining tags for texts of interest in the humanities context, for example <front> (front matter of a book), <castList> (a list of actors in a performance), tags for poetry, etc.
- TEX and LATEX, providing tags to format documents, such as \textbf (for bold font) or $\continuous (for the caption of a figure or table).$

Other document formats, like Microsoft Word, use covert markup, i.e. markup that the user assigns but which then becomes more or less invisible, so that the document appears to the eye to be formatted without any markup.

Whether markup is overt or covert, the final document contains only a visual display which readers have come to understand: paragraphs are separated by spaces and possibly indented; Q-and-A lists are lists of paragraphs starting with a 'Q' or 'A'; lines of poetry use a ragged right margin; section titles are short lines, often in a larger font, and possibly preceded by a section number, and text which is bolded is represented by glyphs whose strokes are thicker.

In general terms, the task we are attempting is to reverse engineer the visual display of a document into an overt markup, by inferring a document's covert structure—the same thing a proficient human reader does when they read a document. We illustrate this problem with a few examples, before turning to some (in-progress) solutions.

Figure 1 shows a small snippet from a PDF document.¹ Two terms are italicized, indicating in this particular document that the terms are being defined. But the italicization shows up

¹The image is taken from a pre-print of Bale and Reiss 2018.

neither in the plain text extracted by the Tesseract OCR program (figure 2), nor in Tesseract's more verbose hOCR output(figure 3).²

and information relating to meaning. We call the information relevant to a morpheme's pronunciation its *phonological representation*, and we call the information relevant to its meaning its *semantic representation*.

Figure 1: Excerpt from PDF, showing italicization

and information relating to meaning. We call the information relevant to a morpheme's pronunciation its phonological representation, and we call the information relevant to its meaning its semantic representation.

Figure 2: Excerpt of plain text OCR output from figure 1

Bold text is similarly not tagged as such in Tesseract's output.³ The implication of this is that the fact that the original author has highlighted something as important, cannot be recovered; structural information has been lost.

Even more problematic is the situation where a textual document represents a highly structured database, and one wishes to reconstruct the structure of the data from the formatting. A common example of this is dictionaries, where lexical information is contained in fields delineated by the formatting. Figure 4 shows an example of this, here a Polish-English dictionary.⁴ Notice that some of the fields within this entry are indicated by switching between bold and nonbold, such as the headword and its definition, or between the sub-entry '**z zapartym tchem**' and its definition "with bated breath"; other fields are indicated by switching between an italic font and a non-italic font, such as '*bez tchu*' and "out of breath."

At a higher level in the document, we also wish to infer things like chapters and sections with their titles, itemized lists, sidebars, footnotes, tables and their captions, and a host of other things that make a document structure apparent to the human reader.

Figure 5 (intentionally shown small to draw attention to the formatting) is a page from National Security Commission on Artificial Intelligence (NSCAI) 2021 (referred to later as the NSCAI report) shows many examples of this. The red text at the top highlights one of four 'Judgments' in the document; a list of these judgments appears in the blue side bar box to the right. A short paragraph of plain black text introduces a bulleted list, with each list item beginning with a bolded word, and using a dark blue font. More plain black text follows the list. A header and footer appear at the top and bottom of the page. Not readily visible in this image are several footnote numbers, indicated by a raised digit in a smaller font.

This document is about 750 pages. If one wanted to know the committee's judgments, a search for simply the word 'judgment' would pull up both the committee's judgments and

²Tesseract reports version 5.1.0-72-gb8b6, with correspondingly updated libraries. It was installed as an Ubuntu-compatible binary from the Tesseract website on 22 July 2022.

³Tesseract version 3, used in M. Maxwell and Bills 2017 and M. Maxwell and Bills 2018, attempted to tag bold text, but its recall and precision were very poor in the documents we worked with. The ability to tag bold text was removed in Tesseract version 4.

⁴The dictionary shown is Oxford University Press 2010.

```
<span class='ocrx_word' id='word_1_15' title='bbox 178 23 194 33;</pre>
   x_wconf 96'>its</span>
  <span class='ocrx_word' id='word_1_16' title='bbox 200 22 277 36;</pre>
   x_wconf 95'>phonological</span>
  <span class='ocrx_word' id='word_1_17' title='bbox 284 23 376 36;</pre>
   x_wconf 96'>representation, </span>
  <span class='ocrx_word' id='word_1_18' title='bbox 383 22 407 33;</pre>
   x_wconf 96'>and</span>
  <span class='ocrx_word' id='word_1_19' title='bbox 413 26 430 33;</pre>
   x_wconf 96'>we</span>
  <span class='ocrx_word' id='word_1_20' title='bbox 435 22 457 33;</pre>
   x_wconf 96'>call</span>
  <span class='ocrx_word' id='word_1_21' title='bbox 463 22 484 33;</pre>
   x_wconf 93'>the</span>
  <span class='ocrx_word' id='word_1_22' title='bbox 490 23 506 33;</pre>
   x_wconf 91'>in-</span>
</span>
<span class='ocr_line' id='line_1_3' title="bbox 2 40 396 54;</pre>
   baseline 0 -3; x_size 17.782608; x_descenders 3;
   x_ascenders 4.782609">
  <span class='ocrx_word' id='word_1_23' title='bbox 2 40 64 51;</pre>
   x_wconf 96'>formation</span>
```

Figure 3: Excerpt from hOCR output for figure 1

all other instances of the word 'judgment', of which there are about 30. Clearly it would be desirable to be able to restrict search to instance of that word in red text, and similarly for other pieces of structurally-delineated information in the report. Structural information based on formatting would also allow one to put together an outline of the document (there is a table of contents, but it includes only the chapter titles).

3 Previous Work

M. Maxwell and Bills (2017; 2018) describes parsing an OCRed Tzeltal-Spanish dictionary (Cruz, Gerdel, and Slocum 1999, similar to the Polish dictionary of figure 4 in its use of bolding)

dech breath: bez tchu out of breath ♦ pozbawiać kogoś tchu wind sb ♦ zapierać ~ get your breath (again/ back) ♦ odczytywać coś jednym tchem reel sth off [IDM] z zapartym tchem with bated breath | zapierać ~ w piersiach (przen.) take your breath away

Figure 4: Entry from Polish-English dictionary

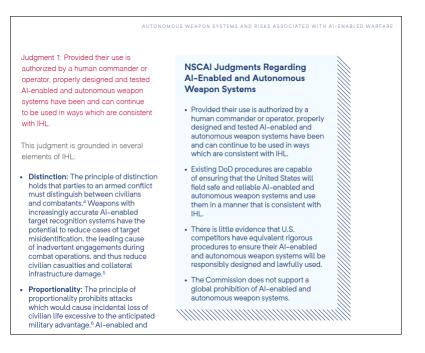


Figure 5: Partial page from NSCAI report

to create an XML database with fields such as headwords, part of speech, definition, sub-entries etc. The failure to detect bold text marking the delineation between the Tzeltal and Spanish portions of subentries resulted in low parsing accuracy.

Document Image Analysis (DIA) tasks include document image classification and layout detection in images; the technology can be used to convert images of documents into structured data. Early work on DIA involved rule-based approaches, but deep learning (neural net) approaches are now more common.

Shen et al. 2021 describes Layout Parser, a trainable deep learning toolkit and model repository for DIA (see also their section on 'Related Work'). They use image processing and OCR to find rectangular boxes of text (and pictures) in document images being processed and label the function of each box; it also provides pre-trained structural models for certain types of documents. But since a given document can have a very different structure from the documents in Layout Parser's existing models, it is desireable to fine-tune a model that is similar to the target documents, or in the worst case to train a new model from scratch. In either case, a set of hand-labeled documents is required for the tuning or training.

Clausner, Pletschacher, and Antonacopoulos 2011 describes the University of Salford's Aletheia, a DIA system including tools for annotating document structure with the assistance of various semi-automated mechanisms. The university also maintains a dataset of annotated documents at https://www.primaresearch.org/dataset/.

In the next section, we describe on-going work in the evaluation and use of DIA tools (including the Layout Parser), while section 5 discusses experiments in improving the detection of bold text.

4 General Format Detection

We experimented with the Layout Parser on various documents, including the NSCAI report (a partial page of which was shown in figure 5). Our goal was to recognize significant text boxes, join boxes that were broken over page or column boundaries, and if possible label the boxes as to type (body text, footers and headers, section titles, etc.).

We selected a pre-trained model that worked reasonably well with our document's format, but even this best model showed some spurious or overlapping boxes in the output. However, we were able to eliminate most of these spurious boxes without losing valid boxes by ignoring boxes to which the program assigned a low confidence; this in turn eliminated most overlapping boxes. In addition, boxes frequently clipped off pieces of characters at the edges, but we addressed this by adding a small padding to increase the size of each box.

The model classified regions as (body) text, titles, lists, tables or figures; it was however confused by sidebar elements (like the one shown in the upper right-hand side of figure 5).

We experimented with the detection of titles by Layout Parser's PubLayNet dataset and its corresponding model (faster_rcnn_R_50_FPN_3x)⁵ by running this over a set of test samples with 15 images. The aim of this experiment was to estimate the quality of the model performance, including a raw count of three categories:

- 1. 'Correct Bold Text' (34 instances): Boxes correctly tagged as titles, where the text was bold.
- 2. 'Incorrect Non-Bold Text' (16 instances): Boxes incorrectly tagged as titles, where the text was not bold.
- 3. 'Incorrect Bold Text' (7 instances): Boxes incorrectly not tagged as titles, where the text was bold.

Figure 6 shows some examples from the test set.

Notice that the raw number of incorrect bold text segments is relatively small compared to the number of bold text segments that are correctly classified. However the number of incorrectly classified non-bold texts is quite high, and another method should be used to lower this number. This model could perhaps be combined with other models or methods to decrease the number of incorrect non-bold text tags (ensemble models), but we did not have time to test this.

5 The Special Problem of Bold Font Detection

As mentioned, Tesseract does not currently attempt to distinguish between bold and non-bold (or italic) text. Since bolding is often used to emphasize important words, as well as distinguishing between fields in semi-structured data (like dictionary entries), we experimented with

⁵https://layout-parser.readthedocs.io/en/latest/notes/modelzoo.html, including the dataset pubLayNet and the model faster_rcnn_R_50_FPN_3x

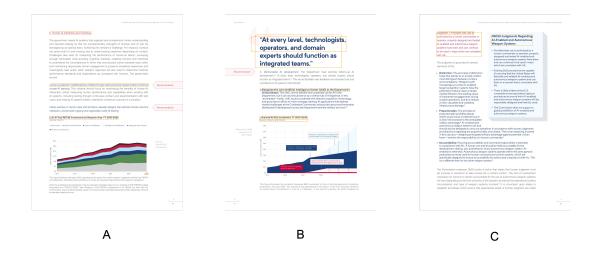


Figure 6: Yellow boxes are marks for titles assigned by the Layout Parser (LP). Picture A shows two non-bold text segments that LP correctly tagged as titles. Picture B shows bold text that LP correctly boxed as titles (although it missed some bold text continued on to the next line). Picture C shows short bold texts at the beginning of the paragraphs that LP did not tag as titles, and red but non-bold text that is incorrectly tagged as a title.

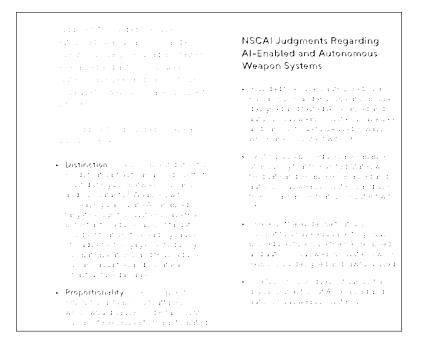
methods for detecting bolded words, using for test data a Tzeltal-Spanish dictionary (Cruz, Gerdel, and Slocum 1999) and a Cubeo-Spanish dictionary (Morse and M. B. Maxwell 1999), as well as the previously mentioned NSCAI document.

The initial attempt used unsupervised clustering based on features such as bounding box dimensions, pixel counts and scaling by letter shape, but this unsupervised approach did not work well.

The best performing method used the OpenCV Python library, an open source library for image processing (https://github.com/opencv/opencv-python). We first converted the color image to gray scale, then 'thresholded' it to convert the gray scale into a black-and-white image (where gray turned into white), and finally used CV2's dilation() function to partially erode areas with black pixels. This had the effect of removing most of the pixels for non-bold characters, while leaving enough pixels in bold characters to enable approximate OCR of the bolded words. The result of applying this to the page shown in figure 5 is illustrated in figure 7; running this image through OCR gives the text shown in figure 8.

Clearly the OCR output from this degraded image is not good enough for downstream use by itself, however by using the bounding boxes in the hOCR output (not shown here), it was possible to pick out the spans in the original hOCR output that contained bold text in the original image.

The CV2 dilation () function has several adjustable parameters. The best settings are doubtless dependent on a number of factors (including the particular font and its size). For our experiments, we set these by hand, but for best results the parameter settings should be adjusted for particular documents or document classes based on a sampling of the outputs.





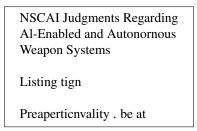


Figure 8: Excerpt of plain text OCR output from figure 7

6 Passing Formatting through Machine Translation

The output of DIA has several potential uses:

- 1. Informing search processes as to which text is more important, which can be used to filter search results.
- 2. Informing summarization processes about important text strings.
- 3. Enabling translations to emulate the formatting of the original text, making the resulting documents more easily understood.

If search (1) or summarization (2) is done in the original document's language, and results are passed through Machine Translation (MT), then the inferred formatting tagging (likely in

XML or JSON) can be stripped before passing the outputs of these processes through MT. However, if search or summarization is instead done on the MT output, or if the MT output is to be formatted according to the input $(3)^6$, then the formatting needs to be passed transparently through the MT process.

How this would be done depends on the particular MT engine; one method would be to train the MT system to ignore such text. Much the same issue arises if named entity tagging is done in the source language, and must be transferred into the target language (cf. Hermjakob, Knight, and Daumé III 2008 for some observations on this problem, albeit in the context of statistical rather than neural net MT). Another method would be to only send blocks of text that are entirely inside a discovered box through translation, and to assemble the translated document using the boxes and their translated contents.

7 Future Work

Clearly much remains to be done before Document Image Analysis can be considered plug-andplay. Too much currently requires manual control: parameter setting (e.g. for the detection of bolding or the choice of confidence-based cutoffs in the boxing of document chunks), choosing pre-trained models, or deciding whether to use an existing pre-trained model as-is or to tune it.

We did not attempt to detect italicization (or underlining, which did not appear in the documents we experimented with). Font size detection is also relevant, e.g. for detecting section headers (which are often in a larger font) or captions or footnotes (often in a smaller font). Font size can probably be inferred from bounding boxes on OCRed text, provided one pays attention to capitalization, and ascenders and descenders.

Finally, some documents may be one-of-a-kind. Figure 9 shows a page from a hand-written Arabic book on Islamic rulings.⁷ The page header includes a 'ruling' title on the left, a page number in the middle, and the 'book' (chapter) title on the right. Below the header is the section title, in red. A horizontal rule about a third of the way down the page separates the original text above from the commentary below. A red font has been used in the commentary to refer to words in bold characters above, as indicated by the orange lines I have drawn in. The words in the commentary circled in green are citations to sources from which the commentator drew, like '*Muraqi al-Falah: page 699*'. The words circled in brown are someone's explanation of the words they appear immediately under. Whether DIA is relevant for such unusual documents probably depends on the size and usefulness of the document in question, and in this case on whether OCR is even possible on this hand-written Arabic document.

In sum, while tools like the Layout Parser require substantial automated training material, for long documents (such as the NSCAI document) or for many documents of a single type (e.g. medical reports or other standardized documents), automating the Document Image Analysis (DIA) process would pay a return on investment by making it easier for users to make sense of Machine Translation output, enabling parsing of structured data such as dictionaries, and informing search by highlighting "hits" that occur in regions of higher importance.

⁶There may also need to be a mapping between input and output formats; for example, when translating italicized text into a script that does not use italicization would require some other form of emphasis.

⁷The image appears at https://theislamshop.com/books/arabic-books/nur-al-idah-arab ic.

أحكام الاعتكاف TTE كتاب الصوم باب الاعتكاف هو الإقامة بنيته في مسجد تقام فيه الجماعة بالفعل للصلوات الخمس، فلا يصح في مسجد لا تقام فيه الجماعة للصلاة على المختار) وللمرأة الاعتكاف في مسجد بيتها وهو محل عينته للصلاة فيه، والاعتكاف على ثلاثة أقسام: واجب في المنذور، وسنة ب فيما سواد، والص كفاية مؤكدة في العشر الأخير من رمضان، ومستحد · فقط وأقله نفلا مدة يسبرة ولو كان ماشيا على المفتى ية كالمدام المس شرعية كالجمعة، أو طبعية كالبول، على المختار) وعن أني يوسف: الاعتكاف الواحب لا يجوز في غير مسجد الجماعة، والنفل يجهز. مرامي الفلاح: ٦٩٩ في مسجد بيتها: لا تخرج منه إذا اعتكفت، فلو خرجت لغير عذر يفسد واجبه، وينتهى نفله، ولو اعتكفت في المسجد فظاهر ما في "النهاية" أنه يكره تنسيزيها، وينبغي على قياس ما صرحوا به من أن المحتار منعهنَّ من الخروج في الصلوات كلها أن لا يتردد في متعهن من الاعتكاف في المسجد (حاشية الطحطاوي: ٦٩٩) وسنة كفاية إلخ: قال الزاهدي: عجبًا للناس كيف تركوا الاعتكاف، وقد كان رسول الله 😹 يفعل الشيء ويتركه، ولم يترك الاعتكاف منذ دخل المدينة إلى أن مات، فهذه المواظبة المقرونة بعدم الترك مرة لما اقترنت بعدم الإنكار على من لم يفعله من الصحابة 🌼، كانت دليل السنية، أي على الكفاية، وإلا كانت دليل الوجوب على الأعيان. [حاشية الطحطاوي: ٧٠٠] فيما سواه: أي في أيَّ وقت شاء سوى العشر الأخير، ولم يكن منذورًا. [مراقى الفلاح: ٧٠١] لصحه المتدور: قلو قال: عليَّ أن أعتكف شهراً بغير صوم، عليه أن يعتكف ويصوم، فإن قيل: لو كان شرطا لكان شرط انعقاد ودوام، وليس كذلك؛ لصحة الشروع فيه ليلا، وكذا يبقى في الليل ولا صوم؟ قلنا: الشرائط إنما تعتبر بحسب الإمكان، ولا إمكان في الليل فيسقط للتعذر، وجعلت الليالي تابعة للأيام كالشرب والطريق في بيع الأرض، ألا ترى أن صلاة المستحاضة تصح مع السيلان، وإن عدم الشرط للتعذر، وكذا الخروج للغائط والبول لا ينافيه للعجز مع أن الركن أقوى من الشرط. (كفاية وحاشية الطحطاوي) يسيرة: غير محددة، فيحصل بمحرد المكث مع النية. [مراقى الفلاح: ٢٠٢] ماشياً: أي ماراً غير حالس في المسحد ولو ليلاً، وهو حيلة من أراد الدخول والخروج من باب آخر في المسجد حتى لا يجعله طريقاً. [مراقى الفلاح: ٢٠٢]

Figure 9: Excerpt from "Nur al-Adiyah"

References

- Bale, A. and C. Reiss (2018). Phonology: A Formal Introduction. MIT Press. ISBN: 9780262348133. URL: https://books.google.com/books?id= t2V0DwAAQBAJ.
- Clausner, C., S. Pletschacher, and A. Antonacopoulos (2011). "Aletheia An Advanced Document Layout and Text Ground-Truthing System for Production Environments." In: *Proceedings of the 11th International Conference on Document Analysis and Recognition* (*ICDAR2011*). Beijing, China, pp. 48–52. URL: https://www.primaresearch. org/www/assets/papers/ICDAR2011_Clausner_Aletheia.pdf.
- Cruz, Manuel A., Florence L. Gerdel, and Marianna C. Slocum (1999). Diccionario tzeltal de Bachajón, Chiapas. Serie de vocabularios y diccionarios indígenas "Mariano Silva y Aceves" 40. Coyoacán, D.F., Mexico: Instituto Lingüístico de Verano, A.C. URL: http://www.sil.org/system/files/reapdata/52/85/76/ 52857610164780871251544555610851968393/S040_DicTzeltalFacs_ tzh.pdf.
- Hermjakob, Ulf, Kevin Knight, and Hal Daumé III (2008). "Name Translation in Statistical Machine Translation — Learning When to Transliterate." In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, pp. 389–397. URL: https: //aclanthology.org/P08-1045.
- Maxwell, Michael and Aric Bills (2017). "Endangered Data for Endangered Languages: Digitizing Print dictionaries." In: Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages. Honolulu: Association for Computational Linguistics, pp. 85–91. URL: http://www.aclweb.org/anthology/ W17-0112.
- (2018). "Giving Digital Life to a Print Dictionary." In: 6th International Conference on Language Documentation and Conservation (ICLDC). Honolulu. URL: https:// scholarspace.manoa.hawaii.edu/bitstreams/45151f4f-d3ba-4782-b5ee-5775f50e430f/download.
- Morse, Nancy L. and Michael B. Maxwell (1999). *Cubeo Grammar*. Studies in the Languages of Colombia 5. Dallas: Summer Institute of Linguistics.
- National Security Commission on Artificial Intelligence (NSCAI) (2021). *Final Report: National Security Commission on Artificial Intelligence*. Arlington, VA: National Security Commission on Artificial Intelligence.
- Norman, Walsh and Richard L. Hamilton (2010). *DocBook 5: The Definitive Guide: The Official Documentation for DocBook*. 1st. Cambridge: O'Reilly.
- Oxford University Press (2010). Oxford Essential Polish Dictionary. Oxford University Press.

Shen, Zejiang, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li (2021). "LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis." In: arXiv preprint arXiv:2103.15348. URL: https://arxiv. org/abs/2103.15348.