An End-to-End Pipeline for Bibliography Extraction from Scientific Articles

Bikash Joshi, Anthi Symeonidou, Syed Mazin Danish, Floris Hermsen
Elsevier

Abstract

We introduce a comprehensive end-to-end pipeline designed to extract complete bibliography section from English scientific articles in digital-born PDF format and further split them into individual citations. At the heart of our pipeline lies the utilization of Language-independent Layout Transformer (LiLT), a multimodal model that combines text and layout features to enhance the accuracy and robustness of bibliography extraction. By considering both text and visual structure, LiLT significantly improves the identification of bibliographic sections within scientific articles. To split the extracted full bibliography into individual citations, we employ a custom fine-tuned version of SciBERT, a Transformer-based model that excels at handling complex formatting variations common in scholarly bibliography.

Having such end-to-end pipeline in-house allows us to bypass reliance on third-party black box tools, such as GROBID, offering greater control and transparency in the bibliography extraction process. Another highlight of our pipeline is its extensibility, as it can be seamlessly adapted to multilingual and image-based PDFs, hence allowing its utility across a wide range of scholarly content. When evaluated on an in-house dataset of digital-born English PDF articles published at Elsevier, we achieved an F1-score of 94.6%, a notable 3.1% improvement over GROBID, which is a well-regarded tool for bibliography parsing in the industry.

1 Introduction

Scientific articles are an essential part of the scientific community. In the digital age, where millions of scientific articles are published every year, efficient extraction of header (title, author names, affiliations, abstract) and bibliography entities from unstructured data, can facilitate not only the searchability and discoverability of scientific work, which is beneficial for the researchers, but it also plays a role in the automation of academic workflows.

Although most scientific articles received by scientific publishers come in semi-structured format (MS Word), a significant proportion of scholarly articles still reside in PDF-based documents. The diverse formatting, layouts, and font styles found in PDF articles demand sophisticated techniques to accurately extract bibliographic information, such as citation details, from these unstructured documents.

By facilitating precise referencing and citation tracking, bibliography extraction aids in the credibility and impact assessment of published research, a critical aspect for publishing companies as they endeavor to maintain the integrity and relevance of the scientific literature they curate. Mature tools such as GROBID (GRO, 2008–2023), Cermine (Tkaczyk et al., 2015) and Neural ParsCit (Prasad et al., 2018a), provide various APIs for header and bibliography entities extraction with good results (Romary and Lopez, 2015; Lo et al., 2020). However, these tools face limitations in coping with scanned documents or multilingual content. Addressing these challenges requires a more tailored and fine-tuned solution.

Most traditional approaches to information extraction from PDF documents have primarily relied on text-based methods as evidenced in (Cioffi and Peroni, 2022; Matsuoka et al., 2016; Prasad et al., 2018b). Document layout analysis with Convolutional Neural Networks (CNNs), visual information extraction with Graph Neural Networks (GNNs) and the emergence of Transformer architecture, have shifted the necessity of many annotated data and improved the accuracy of document layout analysis tasks (Zhong et al., 2019; Qasim et al., 2019). However, with the advent of Document AI, there has been a notable shift towards multimodal approaches that seamlessly integrate both textual and layout features (Cui et al., 2021). One prominent example of such a multimodal approach is LayoutLM, along with its subsequent

Proceedings of the 2nd Workshop on Information Extraction from Scientific Publications, pages 101–106
Nov 1, 2023.

versions, LayoutLMv2 and LayoutLMv3. These models represent pre-trained Document Foundation Models that effectively merge Natural Language Processing (NLP) and Computer Vision (CV) technologies and substantially outperform several text-based SOTA pre-trained models such as BERT and RoBERTa (Xu et al., 2020, 2022; Huang et al., 2022). Li et al also showed that the LayoutLM model shows better detection accuracy on the DocBank, a benchmark dataset for document layout analysis when compared with other transformer-based or R-CNN models (Li et al., 2020). However, the license of the LayoutLMv3 prohibits it from being used in industry. A good alternative for industrial use cases instead, is the Language-independent Layout Transformer (LiLT), a multimodal model, which overcomes the language barrier and decouples and learns the layout knowledge from the monolingual structured documents before generalizing it to the multilingual (Wang et al., 2022).

Our approach focuses on employing a multimodal approach to navigate the complexities of PDF articles and extract bibliographic data with precision, without depending on external tools for which we don't have the ability to alter their behavior, with the additional opportunity to expand to multilingual content.

2 Grobid Pipeline

GeneRation Of Bibliographic Data (GROBID) (GRO, 2008–2023) is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured XML/TEI encoded documents with a particular focus on technical and scientific publications. GROBID provides APIs for extraction of entities from both Head and Tail (bibliography) sections of PDF manuscripts. GROBID is popularly used for entity extraction from scientific articles and serves as a strong baseline for entity extraction from both header and bibliography. This tool has been around for more than a decade and considered a standard tool in both academia and industry (Lipinski et al., 2013).

3 In-house Bibliography Extraction Pipeline

In this work, we developed an in-house pipeline for extracting citations from PDF articles. This pipeline takes PDF articles as input and gives a list of citations as the final output. Figure 2 depicts the details of this pipeline. This pipeline is composed of the following main components:

3.1 PDF Parser

This component enables the extraction of text and layout information from the input PDFs. As shown in Figure 2, we also have a rule-based candidate selection logic, which helps us to select a few candidate pages containing bibliography. We experimented with various tools for parsing the selected PDF pages, two of which seemed particularly promising:

- PyMuPDF[1] is a Python-based PDF parser, which is actively maintained and enhanced with over 30 million downloads. This ease of use makes this tool quite popular across several entity extraction applications.

- PDFlib TET (Text and Image Extraction Toolkit)[2] is a library written in C/C++. It provides bindings for various programming languages, including Python. Also, it provides a binary executable, which can be invoked from various computational environments.

3.2 Bibliography Detector

The next module in our pipeline is the bibliography detection model, which takes the text and layout extracted by the PDF parser as input and performs token classification for each token, classifying them as either bibliography or non-bibliography. As the multimodal token classification model, we use the Language-independent Layout Transformer (LiLT).

LiLT (Wang et al., 2022) is a multimodal model which takes both text and bounding boxes as input. The entire framework represents a parallel dual-stream Transformer that concurrently processes two streams of information: one for text and the other for layout.

LiLT can be pre-trained on the structured documents of a single language and then directly fine-tuned on other languages with the corresponding off-the-shelf monolingual/multilingual pre-trained textual models. This transfer learning enables multimodal document understanding for many languages, potentially very useful in the context of applications that require multilingual capability. The LiLT architecture is shown in Figure 3.

[1] https://pymupdf.readthedocs.io
[2] https://www.pdflib.com/products/tet/

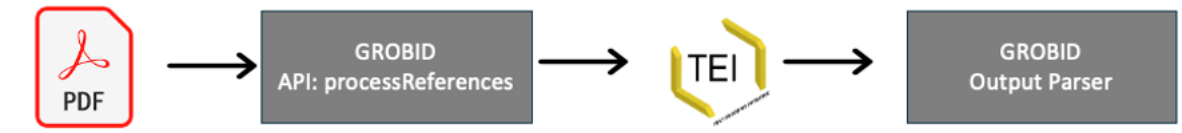

Figure 1: Grobid bibliography extraction pipeline

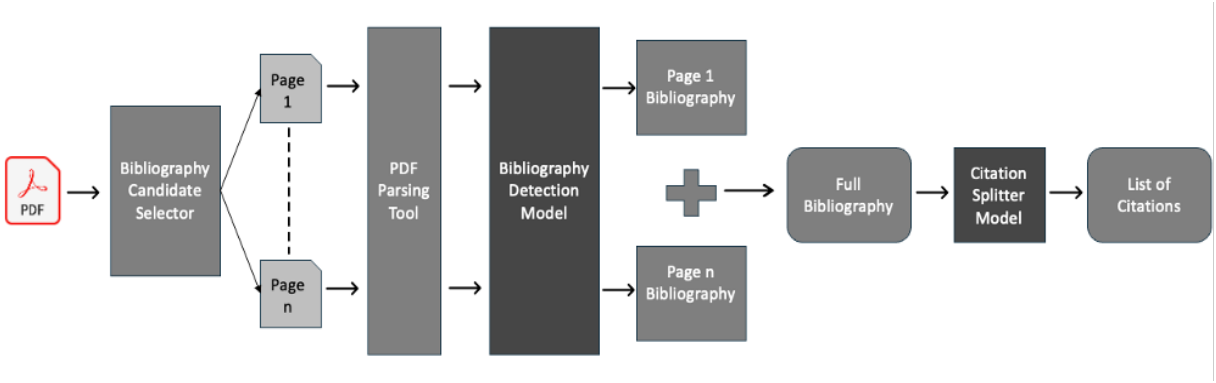

Figure 2: In-house bibliography extraction pipeline

3.3 Citation Splitter

The last component of this pipeline is a citation splitting model, designed to divide the full bibliography into separate citations. While this might seem a straightforward task, it presents a formidable challenge for machine learning algorithms due to the considerable diversity in citation formats.

In this work we fine-tune SciBERT (Beltagy et al., 2019), a BERT-like, transformer-based model trained on scientific content. We trained this customized SciBERT model as a token classifier, employing an in-house dataset of bibliographies for supervised learning. This approach enabled the model to learn to accurately detect the starting point of each citation within the bibliography. As new citations consistently commence after a newline in scientific articles, we made an additional effort to simplify the task for the model by retaining the newline information within the complete bibliography text as an extra clue for the model.

4 Experiments

4.1 Datasets

For training the bibliography detection model, we conducted experiments using two publicly available datasets: DocBank (Li et al., 2020) and GROTOAP2 (Tkaczyk et al., 2014). Our preliminary analysis and experimentation demonstrated the superiority of GROTOAP2 dataset over DocBank dataset in terms of its annotation quality.

To train the citation splitter model, we used an in-house dataset of bibliographies, by annotating the starting point of each citation within the bibliography.

For the final evaluation, we used scientific PDF articles in English from Elsevier's internal scientific articles database published after 2020. All experimental results reported in this article were conducted on this in-house dataset.

4.2 Compared Methods

We compare the following approaches:

- GROBID-CRF: GROBID with CRF-based models.
- GROBID-DL: GROBID with Deep Learning based models. As recommended in the documentation, we use BiLSTM-CRF model.
- In-house pipelines: A proposed stack of in-house models, with PyMuPDF and PDFlib as PDF parsing tools, LiLT as a bibliography detection model and a SciBERT-based citation splitter model.

4.3 Experimental Results

Table 1 shows the final results obtained in our experiments. We evaluated the extraction of the full bibliography section and the extraction of each citation in the bibliography.

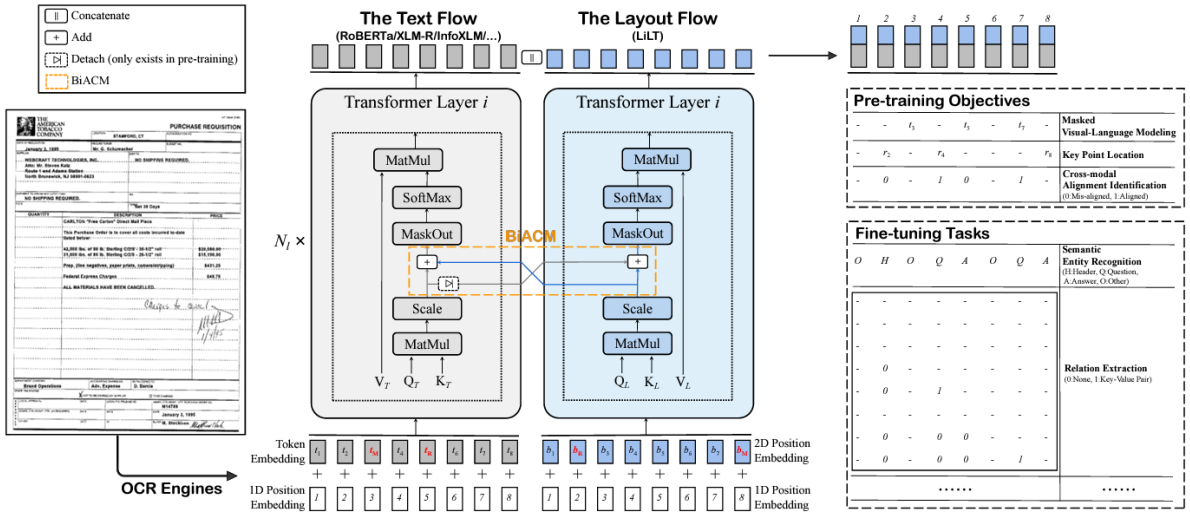

Figure 3: Language-Independent Layout Transformer (LiLT)

Pipeline	PDF Parsing Tool	Bibliography	Citations		
		Accuracy	Precision	Recall	F1
GROBID-CRF	pdfalto	50.2	90.8	89.8	90.3
GROBID-DL	pdfalto	50.9	92.2	90.8	91.5
In-house	PyMuPDF	66.7	92.4	93.2	92.8
In-house	PDFlib	**72.8**	**94.9**	**94.3**	**94.6**

Table 1: Comparison of different methods for extracting bibliography and citations

Full bibliography detection is evaluated in terms of accuracy, calculated as the ratio of the number of correctly detected bibliographies to the total number of bibliographies present in the evaluation dataset (GRO, 2008–2023; Ohta et al., 2014). The correctness of the bibliography is measured in terms of a relaxed edit distance (Levenshtein), keeping a tolerance of up to 10 consecutive mistakes and 10% total mistakes in terms of normalized edit distance). Evaluation at the citation-level is performed in terms of precision, recall and F1-score (GRO, 2008–2023). The metrics are defined as follows:

- Precision: Ratio of the number of correctly extracted citations to the total number of citations extracted by the system.
- Recall: Ratio of the number of correctly extracted citations to the number of all citations in the ground truth.
- F1 Score: 2* Precision * Recall / (Precision + Recall)

In the experimental results, we observed that GROBID is a strong baseline for this task. It is a very robust tool for performing entity extraction from PDF articles, especially the bibliography. Out of the two variants of GROBID, Grobid-DL was found superior. The proposed in-house pipeline with PDFlib PDF parser, LiLT based bibliography detection model and SciBERT based citation splitter is the best performing pipeline, outperforming Grobid baseline by a large margin of 3.1%. Among the two PDF parsing tools, PDFlib resulted in superior performance especially in terms of reading order detection and extraction of line and paragraph level information, which further allowed us to correct some of the prediction mistakes made at the token level.

5 Conclusion and Future Work

We have presented an end-to-end pipeline for the extraction of bibliographic information from scientific articles in digital-born PDF format. Our pipeline is designed to address the challenges posed by the diverse formatting, layouts, and font styles found in PDF articles. We have leveraged cutting-edge techniques and models, including LiLT and SciBERT, to achieve accurate and robust bibliography extraction. We achieved a significant improve-

ment in accuracy over existing tools like GROBID, showcasing the potential of our approach in advancing the task of bibliography parsing.

We see several avenues for future research. One potential direction would be to integrate generative AI based Large Language Models (LLM) into the pipeline. The versatility of LLMs would increase the adaptability of our pipeline to a wider range of scholarly content, encompassing diverse research domains, languages, and publication formats. Alternatively, our LiLT-based pipeline could be adapted to handle languages other than English through transfer learning, which would be valuable as scientific research is conducted globally.

References

2008–2023. Grobid. https://github.com/kermitt2/grobid.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Alessia Cioffi and Silvio Peroni. 2022. Structured references from pdf articles: assessing the tools for bibliographic reference extraction and parsing.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications.

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

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. Docbank: A benchmark dataset for document layout analysis. *arXiv preprint arXiv:2006.01038*.

Mario Lipinski, Kevin Yao, Corinna Breitinger, Joeran Beel, and Bela Gipp. 2013. Evaluation of header metadata extraction approaches and tools for scientific pdf documents. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 385–386.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. S2orc: The semantic scholar open research corpus.

Daiki Matsuoka, Manabu Ohta, Atsuhiro Takasu, and Jun Adachi. 2016. Examination of effective features for crf-based bibliography extraction from reference strings. In *2016 eleventh international conference on digital information management (ICDIM)*, pages 243–248. IEEE.

Manabu Ohta, Daiki Arauchi, Atsuhiro Takasu, and Jun Adachi. 2014. Empirical evaluation of crf-based bibliography extraction from reference strings. pages 287–292.

Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018a. Neural parscit: A deep learning based reference string parser. *International Journal on Digital Libraries*, 19:323–337.

Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018b. Neural parscit: a deep learning-based reference string parser. *International journal on digital libraries*, 19:323–337.

Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. 2019. Rethinking table recognition using graph neural networks. pages 142–147.

Laurent Romary and Patrice Lopez. 2015. Grobid-information extraction from scientific publications. *ERCIM News*, 100.

Dominika Tkaczyk, Pawel Szostek, and Lukasz Bolikowski. 2014. Grotoap2-the methodology of creating a large ground truth dataset of scientific articles. *D-Lib Magazine*, 20(11/12).

Dominika Tkaczyk, Pawel Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Lukasz Bolikowski. 2015. Cermine: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2022. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. ACM.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. Publaynet: largest dataset ever for document layout analysis.