ETDSuite: An Library for Mining Electronic Theses and Dissertations

Muntabir Hasan Choudhury
Old Dominion University
Norfolk, Virginia, USA
mchou001@odu.edu

Abstract

With the growing interest in studying scholarly data, one of the understudied types of scholarly data – Electronic Theses and Dissertations (ETDs) needs more attention as ETDs have distinct features compared with conference proceedings and journal articles in many aspects. They are book-length documents, the topics may shift across chapters, exhibits the major contribution of the research work of a student, and have different metadata schema (e.g., university, department, disciplines) from regular scholarly papers. Most existing frameworks are designed for journals and conference proceedings. There is a lack of frameworks to extract information from ETDs, including ETD segmentation, metadata extraction, metadata quality improvement, and parsing reference strings. To address the gap, we develop ETDSuite, a library that consists of various frameworks for mining ETDs by exploiting artificial intelligence (AI) methods. We demonstrate the performance of the frameworks at the preliminary stage and propose tasks to improve the performance of existing modules and add modules with new functions.

Keywords

Digital Libraries, ETD, Machine Learning, Deep Learning, Natural Language Processing, Computer Vision

ACM Reference Format:
Muntabir Hasan Choudhury. 2023. ETDSuite: An Library for Mining Electronic Theses and Dissertations. In *Proceedings of 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL'23).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

ETDs represent scholarly works of students pursuing higher education and successfully meeting the partial requirement of academic degrees. Since 1997, pioneered by Virginia Tech, students have been required to submit their theses and dissertations electronically (i.e., ETDs) hosted by university digital libraries (DLs). ETDs contain rich metadata, bibliographies, figures, tables, and discoveries in specific subject areas. However, DL still lacks computation models, tools, and services for discovering and accessing the knowledge found in ETDs. To build a scalable DL, one can exploit machine learning and deep learning methods to extract and parse key document elements of these ETDs to improve the accessibility and discoverability problems, allowing users to easily assess their relevance.

To address the limitations, Ahuja et al. proposed an object detection model by fine-tuning YOLOv7 [20] on the ETD dataset to classify several document elements [1]. This method is powerful for automatically annotating a fraction of major structural components but still underperformed in detecting minority classes (e.g., date, degree, equation, algorithms) due to a lack of training samples. LayoutLMv2 [22] was introduced to perform downstream document understanding tasks (e.g., entity extraction and document image classification) and evaluated against different evaluation benchmarks. For example, the RVL-CDIP dataset [9] was used for evaluating document image classification tasks, consisting of scanned document images belonging to 16 classes (e.g., letter, form, email, resume). We fine-tuned LayoutLMv2 [22] on ETDs for segmenting pages, but it performed poorly (e.g., achieved 9% accuracy). In addition, few other SOTA task-specific applications (e.g., GROBID [14]) can parse bibliographic data published in journals and conference proceedings. However, these applications are not able to extract metadata from ETD cover pages. Moreover, the DL of ETDs is accompanied by incomplete, inconsistent, and incorrect metadata. To our best knowledge, no AI-based frameworks have been proposed and implemented to improve the metadata quality for ETDs. Hence, further methods must be addressed to mine ETDs because ETDs have complex document structures, and low-resolution scanned images of typewritten and handwritten text make OCR-ing (i.e., Optical Character Recognition task) challenging.

Therefore, our main contribution to this thesis is to propose a library called ETDSuite, containing novel methods that segment, extract, parse, and restructure raw ETD documents into structured JSON documents leveraging natural language processing (NLP) and computer vision (CV) models. As a part of the preliminary work, we introduce AutoMeta [10] and MetaEnhance [4] to extract metadata from ETDs and improve the metadata quality in the DL of ETD repositories. We will further demonstrate our ongoing research effort on segmenting and parsing citations of ETDs.

2 Research Questions

The research questions (RQs) are the following:

- **RQ 1:** Segmenting ETDs will allow us to perform ETD mining tasks. Can we develop a multimodal framework that classifies ETD pages by different types?
- **RQ 2:** ETDs can be scanned and born-digital (e.g., using LaTeX). Can we build an AI method to extract metadata from ETDs?
- **RQ 3:** Library provided metadata often exhibit incomplete, inconsistent, and incorrect values. How can we leverage AI methods instead of a manual effort [21] to improve metadata quality?

JCDL'23, June 26–30, 2023, Santa Fe, New Mexico

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

- **RQ 4:** Parsing citations allow us to enrich metadata by linking ETDs to millions of papers in a public DL corpus [13]. Can we develop a method to parse citations in many styles from ETDs?

3 Data Collection

We crawled more than 114 US university libraries, collected over 533,047 ETDs, including full text in PDF format, and harvested metadata from university library repositories and ProQuest [18]. The major datasets we have built based on this collection include the following: a) **ETD500** – it consists of 500 annotated metadata of ETD cover pages in PDF, TXT, XML, and TIFF format, including library-provided ground truth metadata. The ETDs in ETD500 were published between 1945 and 1990. There are 350 STEM and 150 non-STEM majors from 468 doctoral, 27 master's, and 5 bachelor's degrees. In addition, we manually annotated 92,375 pages of ETD500 into 14 categories (e.g., abstract, chapters, dedication), available in PNG format, from which we extracted text and the bounding boxes (bbox) using AWS Textract. b) **ETDQual500** – it consists of 500 ETD benchmark evaluations (different from ETD500) by combining subsets (i.e., 4 ETD subsets from university, year, department, and degree fields) sampled using multiple criteria. The selection criteria ensure that we cover samples with errors (e.g., missing values, acronyms, misspellings) in different metadata fields.

4 Methodology

4.1 Preliminary Work

Metadata Extraction To overcome the limitation that existing frameworks such as GROBID [14] do not work well on parsing ETDs, we first introduced a baseline method [5] using regular expressions by analyzing the text patterns to extract metadata. We applied Tesseract-OCR to extract text and then applied the rule-based method. The method achieved up to 97% accuracy depending on the metadata fields. However, this method performed poorly on ETD500 as it was biased and trained on only 100 ETD cover pages from MIT and Virginia Tech. To address the limitation, we implemented AutoMeta [10]. It uses conditional random field (CRF) [17], a sequence classifier that leverages text and visual features. The model was evaluated on ETD500 and achieved an F1 score ranging from 81.3% – 96% depending on the metadata fields.

Metadata Quality Improvement Addressing the metadata quality problem (e.g., missing values, incomplete, and inconsistent) found in DLs of ETDs, we implemented MetaEnhance [4], used AI models and achieved a remarkable performance against ETDQual500 in detecting, correcting, and canonicalizing metadata errors, achieved an F1 score of 0.85 - 1.00 depending on the metadata fields.

4.2 Propose Work

Multimodal ETD Segmentation The existing SOTA models (e.g., DocFormer [2]) for document understanding task uses the multimodal model that employs visual modalities (i.e., uses RESNET50 [11] or RCNN [7]) and text modalities (i.e., uses transformers [19]) with an attention mechanism by fusing visual, text, and spatial features (e.g., bbox using Tesseract-OCR). Despite the novelty of the architecture, these models [2] performed poorly on ETDs (e.g., DocFormer performs 26% F1 score). Due to this limitation, using ETD500, we propose a multimodal model that uses a vision encoder (e.g., RCNN [7] with bbox) and a text encoder (e.g., LMs [6] [12]) to extract individual embeddings. Later, we will combine the embeddings in an identical space. Further, we will apply the multimodal model to segment ETDs. To achieve a better performance, we will adopt Faster-RCNN [16] for the vision encoder and introduce CRF [17] at the softmax layer to segment ETD pages.

Citation Parsing Academic disciplines have adopted different citation styles in their research. For example, the APA format is commonly used in Education, while the MLA format is used in Humanities. Due to the variation of citation styles, automatically parsing citations became challenging. Existing frameworks such as Neural ParsCit [15]) use deep learning to overcome the challenge of parsing citations accurately but are trained on focused domains with fewer citation styles. To address the limitation, we will fine-tune BERT [6] on citation strings using the GIANT-1B [8] dataset, containing synthesized annotated citations with 1500 styles. Further, we will use a sequence classifier such as CRF [17] as a decoder to parse citations. To evaluate the performance, we will build an evaluation benchmark with 1000 citation strings in many styles from ETDs focusing on key metadata fields (e.g., title, author, venue, and year). Parsing citations from ETDs will help us predict the career trajectories of graduate students [3] by building a citation network.

Enhancing Metadata Quality AutoMeta [10] can extract metadata from the first page of ETD. We found this cover page expands to more than one page, containing a list of advisor names appearing on the following page other than the first page. To overcome this limitation, integrating the ETD segmentation framework will help AutoMeta [10] to be more scalable and reliable in extracting metadata fields from more than the first page. Moreover, MetaEnhance [4] uses AutoMeta [10] to fill in missing values and overwrite the incorrect values. However, AutoMeta [10] achieved F1 scores of 0.67 to 0.91 against ETDQual500 for most metadata fields. Therefore, enhancing the ability of AutoMeta [10] will significantly help MetaEnhance [4] to achieve the best performance. In addition, despite the performance of MetaEnhance [4] against ETDQual500, the dataset lacks a significant amount of data points in two of four major criteria, including misspellings and incorrect values to measure the robustness of the model. Hence, we will build a challenging dataset by augmenting more misspellings and incorrect surface values to the ETDQual500.

5 Conclusion

We introduced ETDSuite, a library that is capable of performing various tasks using machine learning and deep learning-based methods by incorporating text and visual features to segment, extract, and parse ETDs. We demonstrated the results in the preliminary works and briefly discussed our ongoing research effort to successfully built such an intelligent system. Moreover, we demonstrated various challenges and the research gap for each framework. Developing ETDSuite will help us perform further downstream tasks. For example, parsing citations from ETDs will help us build a citation recommendation system for ETDs. Thus, building the ETDSuite library will benefit all librarians, students, and scientists in the academia and digital library domain.

References

[1] Aman Ahuja, Alan Devera, and Edward Alan Fox. 2022. Parsing Electronic Theses and Dissertations Using Object Detection. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*. Association for Computational Linguistics, Online, 121–130. https://aclanthology.org/2022.wiesp-1.14

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. DocFormer: End-to-End Transformer for Document Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 993–1003.

[3] Clara Boothby, Stasa Milojevic, Vincent larivière, Filippo Radicchi, and Cassidy Sugimoto. 2022. Consistent churn of early career researchers: an analysis of turnover and replacement in the scientific workforce. https://doi.org/10.31219/osf.io/hdny6

[4] Muntabir Hasan Choudhury, Lamia Salsabil, Himarsha R. Jayanetti, Jian Wu, William A. Ingram, and Edward A. Fox. 2023. MetaEnhance: Metadata Quality Improvement for Electronic Theses and Dissertations of University Libraries. arXiv:2303.17661 [cs.DL]

[5] Muntabir Hasan Choudhury, Jian Wu, William A. Ingram, and Edward A. Fox. 2020. A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*. ACM, 515–516. https://doi.org/10.1145/3383583.3398590

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 580–587. https://doi.org/10.1109/CVPR.2014.81

[8] Mark Grennan, Martin Schibel, Andrew Collins, and Joeran Beel. 2019. GIANT: The 1-Billion Annotated Synthetic Bibliographic-Reference-String Dataset for Deep Citation Parsing. In *Irish Conference on Artificial Intelligence and Cognitive Science*.

[9] Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. 991–995. https://doi.org/10.1109/ICDAR.2015.7333910

[10] Muntabir Hasan Choudhury, Himarsha R. Jayanetti, Jian Wu, William A. Ingram, and Edward A. Fox. 2021. Automatic Metadata Extraction Incorporating Visual Features from Scanned Electronic Theses and Dissertations. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 230–233. https://doi.org/10.1109/JCDL52503.2021.00066

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692 (2019).

[13] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[14] Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries, 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Proceedings (Lecture Notes in Computer Science, Vol. 5714)*. Springer, 473–474. https://doi.org/10.1007/978-3-642-04346-8_62

[15] Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018. Neural ParsCit: A Deep Learning Based Reference String Parser. *International Journal on Digital Libraries* 19 (2018), 323–337. https://link.springer.com/article/10.1007/s00799-018-0242-1

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[17] Charles Sutton and Andrew McCallum. 2012. An Introduction to Conditional Random Fields. *Found. Trends Mach. Learn.* 4, 4 (apr 2012), 267–373. https://doi.org/10.1561/2200000013

[18] Sami Uddin, Bipasha Banerjee, Jian Wu, William A. Ingram, and Edward A. Fox. 2021. Building A Large Collection of Multi-domain Electronic Theses and Dissertations. In *2021 IEEE International Conference on Big Data (Big Data)*. 6043–6045. https://doi.org/10.1109/BigData52589.2021.9672058

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[20] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696 [cs.CV]

[21] Jian Wu, Kyle Williams, Madian Khabsa, and C. Lee Giles. 2014. The impact of user corrections on a crawl-based digital library: A CiteSeerX perspective. In *10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*. 171–176.

[22] 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. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201