Automatic Metadata Extraction Incorporating Visual Features from Scanned Electronic Theses and Dissertations

1st Muntabir Hasan Choudhury
Department of Computer Science
Old Dominion University
Norfolk, United States
mchou001@odu.edu

2nd Himarsha R. Jayanetti
3rd Jian Wu
Department of Computer Science
Old Dominion University
Norfolk, United States
{hjaya002, j1wu}@odu.edu

4th William A. Ingram
5th Edward A. Fox
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, United States
{waingram, fox}@vt.edu

***Abstract*—Electronic Theses and Dissertations (ETDs) contain domain knowledge that can be used for many digital library tasks, such as analyzing citation networks and predicting research trends. Automatic metadata extraction is important to build scalable digital library search engines. Most existing methods are designed for born-digital documents such as GROBID, CERMINE, and ParsCit, so they often fail to extract metadata from scanned documents such as for ETDs. Traditional sequence tagging methods mainly rely on text-based features. In this paper, we propose a conditional random field (CRF) model that combines text-based and visual features. To verify the robustness of our model, we extended an existing corpus and created a new ground truth corpus consisting of 500 ETD cover pages with human validated metadata. Our experiments show that CRF with visual features outperformed both a heuristic baseline and a CRF model with only text-based features. The proposed model achieved 81.3%-96% F1 measure on seven metadata fields. The data and source code are publicly available on Google Drive[1] and a GitHub repository[2], respectively.**

***Index Terms*—Digital Libraries, Optical Character Recognition, Text Mining, Metadata Extraction, CRF, BiLSTM**

I. Introduction

A thesis or dissertation is one type of scholarly work that shows a student pursuing higher education has successfully met key requirements of a degree. An ETD is usually accessible from a university's digital library or a third-party ETD repository such as ProQuest. Since the inception of ETDs around 1997, pioneered by Virginia Tech, many ETDs are generated electronically (i.e., born-digital) by computer software such as LaTeX and Microsoft Word. However, the majority of the ETDs produced before 1997 and a significant fraction of ETDs after 1997 are scanned from physical copies (i.e., non-born-digital). These ETDs are valuable for digital preservation, but to make them accessible, it is necessary to index the metadata of these ETDs.

Many ETD repositories are accompanied by incomplete, little, or no metadata, posing challenges for accessibility. For example, advisor names appearing on the scanned ETDs may not be available in the metadata provided in the library repository. Thus, an automatic approach should be adopted to extract metadata from scanned ETDs. Several tools [1]–[4] have been developed to automatically extract metadata for relatively short and born-digital documents, such as conference proceedings and journal articles published in recent years. However, they do not work well with scanned book-length documents such as ETDs. Extracting metadata from scanned ETDs is challenging due to poor image resolution, typewritten text, and imperfections of the OCR technology. The first step is to extract text from PDF files using the Optical Character Recognition (OCR) technique. Many commercial-based OCR tools such as OmniPage, ABBYY FineReader, or Google Cloud API OCR could be used for converting PDFs to text, but they usually have a cost. We adopted Tesseract-OCR, an open-source OCR tool, to extract text from the cover pages of scanned ETDs. Tesseract-OCR supports printed and scanned documents and more than 100 languages. It returns output in plain text, hOCR, PDF, and other formats.

In the preliminary work [5], we proposed a heuristic method to extract metadata from the cover pages of scanned ETDs. However, heuristic methods usually do not generalize well. They often fail when applied to new data with a different feature distribution. In this paper, we investigate the possibility of improving the generalizability of our method using a learning-based model.

II. Related Work

Several frameworks have been proposed to extract metadata from scholarly papers. CERMINE [4] was developed to extract structured bibliographic data from scientific articles. It can extract information related to title, author, author's affiliation, abstract, keywords, journal, volume, issue, pages, and year. For the metadata extraction tasks, they used both machine learning models such as Support Vector Machine (SVM) and simple rule based models. CERMINE achieved an average F1 score of 77.5% for most metadata types and the benchmark

[1]https://tinyurl.com/y8kxzwrp
[2]https://github.com/lamps-lab/ETDMiner/tree/master/etd_crf

978-1-6654-1770-9/21/$31.00 ©2021 IEEE
DOI 10.1109/JCDL52503.2021.00066

evaluation outperformed the existing tools, including GROBID [3] and ParsCit [6], while extracting metadata such as title, email addresses, year, and references. One limitation of this tool is that the PDF documents which contain scanned pages will not be properly processed.

GROBID [3] is a text mining library for extracting bibliographic metadata from born-digital papers. GROBID is based on eleven different CRF models and each parses text using position (e.g., beginning or ending of the line), lexical information, and layout information. It is capable of extracting header and bibliographic metadata such as title, authors, affiliations, abstract, date, keywords, and references. It achieves an accuracy of 74.9% per complete header instance on the CORA dataset but it fails to extract metadata from non-born-digital documents such as scanned ETDs.

In our previous work [5], we have introduced a heuristic model to extract metadata fields from scanned ETD cover pages. It is a rule based method where the metadata fields are captured using a set of carefully designed regular expressions.

III. Dataset

The dataset used in our previous study [5] consisted of a relatively small number of ETDs from only two universities. To overcome this limitation, we created a new dataset of 500 ETDs, which includes 450 ETDs from 15 US universities and 50 ETDs from 6 non-US universities as illustrated in Figure 1. These ETDs were published between 1945 and 1990. There are 350 STEM (Science, Technology, Engineering, and Mathematics) and 150 non-STEM majors from 468 doctoral, 27 master's, and 5 bachelor's degrees. We derived the following seven intermediate datasets that are generated in different stages in our pipeline.

1) The cover page of each ETD in PDF format.
2) TIFF images of (1). The TIFF format is used as the input to Tesseract because it tends to produce fewer errors than JPEG.
3) TXT-OCR: The output of Tesseract containing noisy text extracted from the TIFF images.
4) TXT-clean: The cleaned version of the TXT-OCR dataset after manually correcting misspellings, fixing OCR mistakes, lowercasing the text, and removing empty lines between text. We did not remove line breaks.
5) TXT-annotated: Seven metadata fields annotated using the GATE annotation tool [7].
6) GT-meta: The ground truth from metadata provided by libraries. The data were gathered in the XML-format from MIT, JSON-format from Virginia Tech, and in HTML format for other universities from ProQuest.
7) GT-rev: Revised metadata from GT-meta after manually rectifying discrepancies between library provided metadata and the data present in the cover page of PDF documents.

We observed several challenges to convert scanned ETDs to text (Figure 2).

1) Not all fields are always available on the cover pages.

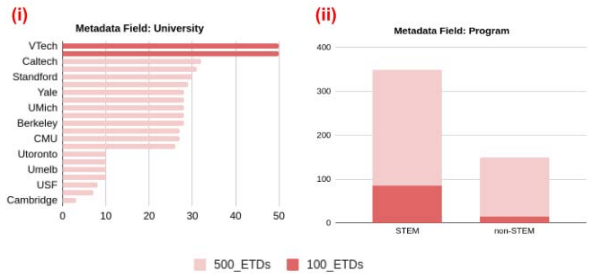


Fig. 1. Distribution of metadata fields: University (i) and Program (ii) in the corpus of 500 ETDs.

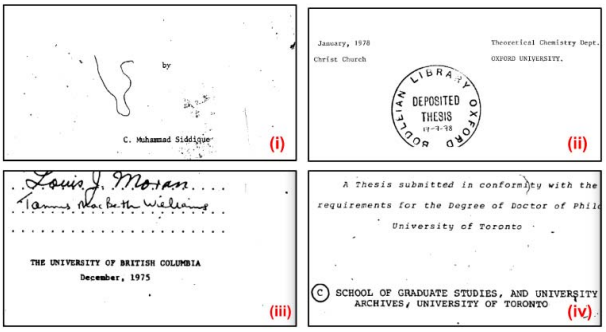


Fig. 2. OCR challenges for ETDs: scribble (i), stamp (ii), overlapped letters (iii), and copyright character (iv).

2) Lines were present to fill the title, degree, author, etc.
3) Multiple years may be provided, such as "submitted year" and "publication year."
4) There were ETD cover pages where author's previous educational certifications are listed making it difficult to extract the degree field.
5) College name (e.g., St. Catherine's College) is mentioned instead of university name (e.g., University of Oxford).

IV. Methodology

A. CRF with Sequence Labeling (CRF Model)

CRF is a statistical modeling algorithm. It differs from classifiers which predict a label for a single sample without considering nearby samples. CRF takes context into account, i.e., that predictions are dependent on each other. We adopted the commonly used BIO (begin, inside, outside) schema. For example, a token can begin an author name, be an inside part of the name, or not be part of that metadata field. The BIO tagging schema has been applied in studies such as named entity recognition [8] [6] and keyphrase extraction [9].

We tagged each token with Part of Speech (POS). POS tags are important here if the phrase consists of pronoun, preposition, or determiner, e.g., "at the Massachusetts Institute of Technology." If the current token "Massachusetts" is tagged as an NNP and the previous token "the" is tagged as a DT, this transition can be used for the model to classify "Massachusetts" as part of a university's name. Our model extracts the following features.

1) Whether all the characters in the word are uppercase.

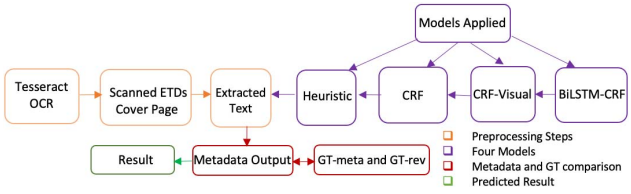


Fig. 3. Metadata Extraction System

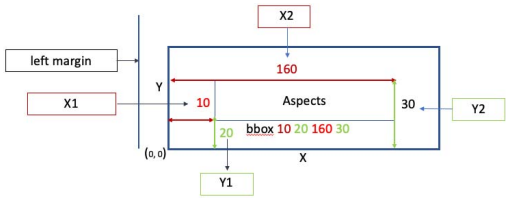
2) Whether all the characters in the word are lowercase.
3) Whether all the characters in a word are numeric.
4) Last three characters of the current word.
5) Last two characters of the current word.
6) POS tag of the current word.
7) Last two characters in the POS tag of the current word.
8) POS of the two tokens after the current word.
9) Whether the first character of consecutive words is uppercase. For example, a title may contain consecutive words that start with an uppercase letter.
10) Whether the first character is uppercase for the word that is not at the beginning or end of the document. This is useful for metadata fields such as author, advisor, program, degree, and university. These fields are not generally at the beginning or end of the document.

B. CRF with Visual Features (CRF-visual)

In the heuristic and the CRF models, we only incorporate text-based features. When humans annotate the documents, they do not rely on text, but visual features, such as the positions of the text and their lengths. For example, thesis titles usually appear in the upper half of the cover page and the authors and advisors usually appear in the lower half of the cover page. This inspires us to investigate whether incorporating visual features can improve performance.

Visual information is represented by corner coordinates of the bounding box (bbox) of a text span. We extract all x-coordinate values (e.g., x1, x2) and y-coordinate values (e.g, y1, y2) for each token. This information is available from hOCR files and XML files, which are output from Tesseract. Figure 4(a) illustrates the bounding box information of the token with x and y coordinates. x1 is the distance from the left margin to the bottom right corner of bbox. y1 is the distance from the bottom margin to the bottom right corner of bbox. x2 is the distance from the left margin to the upper right corner of bbox. y2 is the distance from the bottom margin to the upper right corner of the bbox. All coordinates are measured with respect to the bottom-left corner of the token.

However, transferring these visual features is challenging because the ground truth text is output by Tesseract and rectified by humans. Therefore, the characters in the rectified text are not necessarily aligned with Tesseract's output. The position information was only available for the OCR output. We applied text alignment using the longest common sequence [10]. In bioinformatics, sequence alignment has been commonly applied to align protein, DNA, and RNA sequences which are usually represented by a string of characters. We used an open-source tool known as Edlib [11] to align the



(a)

```
Submit-ed ---department of Mechanical sngineering
||||||-|||---||||||||||||||||||||||||.||||||||||
Submitted to department of Mechanical Engineering
```

(b)

Fig. 4. (a) Bounding box measurement (b) OCR output text (i.e., noisy) alignment with clean text

TABLE I
A COMPARISON OF THE HEURISTIC MODEL BETWEEN TWO DATASETS. NUMBERS IN PARENTHESES ARE SAMPLE SIZES.

Field	Accuracy% (100)	Accuracy% (500)
Title	81%	45.0%
Author	78%	62.8%
Degree	81%	58.0%
Program	97%	8.0%
Institution	94%	18.8%
Year	65%	37.8%
Advisor	36%	5.0%

noisy text data and clean text. Edlib computes the similarity and minimum edit distance between two text sequences. Then we map the positions for each token from TXT-OCR to TXT-clean. Figure 4(b) illustrates an example of the alignment. We incorporated three visual features, each normalized between 0 and 1, to enhance the performance of the CRF model: (1) Left margin (x1) for all tokens in the same line; (2) Upper left corner position (y2) for all tokens; and (3) Bottom right corner position (y1) for all tokens.

C. BiLSTM-CRF Model

Bidirectional Long Short Term Memory (BiLSTM) has been proven to be effective in sequence labeling problems [12]. BiLSTM learns the context of the given sentence in both forward and backward directions. The architecture of the classifier consists of three layers: word-embedding, BiLSTM as encoder, and CRF. The BiLSTM layer learns the forward and backward context in a sequence and feeds it into the CRF layer, which classifies tokens based on their encodings. We used Adam as the optimizer and Keras word embedding initialized with random weights. The batch size is set to 32, and the model runs for up to ten epochs.

V. EVALUATION AND RESULTS

A. Heuristic Model

In a previous study we applied the heuristic method to 100 ETDs [5]. As shown in Table I, when we applied the method to our new dataset of 500 ETDs, the accuracy was considerably

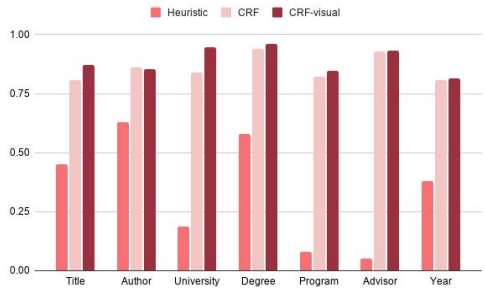


Fig. 5. Performance (F1) comparison of the models.

lower. This is because the new dataset contains ETDs from a more diverse set of universities and majors.

B. CRF Model and CRF-Visual

We randomly divide the samples into two sets: 350 samples for training and 150 samples for testing. The CRF model predicts the labels for each metadata field at the token level. However, we must glue together the predicted tokens for each metadata field and compare them against the ground truth datasets GT-meta and GT-rev. When comparing the title field, some predicted titles did not match exactly, with a small difference such as a punctuation mark or a space character. For example, the model predicted the title as "thermo- fluid dynamics of separated two - phase flow." However, in the GT meta it is "thermo-fluid dynamics of separated two-phase flow." These small offsets are not caused by the model but by line breaks and additional punctuation marks added in text justification. Therefore, the predicted span should be counted as a true positive. We use a fuzzy matching algorithm based on Levenshtein distance and set a threshold of 0.95 when matching predicted and ground truth titles. Figure 5 illustrates the performance of our model. The model outperformed the baseline model whereas CRF-visual outperformed both the baseline model and CRF.

C. BiLSTM-CRF Model

The BiLSTM-CRF model generated poor results for all fields. The F1 scores for token level labels such as B-title, I-title, B-author, and I-author were only 34%, 34%, 24%, and 23%, respectively. The F1 measures for the remaining fields were even lower, so we did not plot them in Figure 5. There are several possible reasons. One major reason is the small size of the training data. The training set contains 350 ETD cover pages. Some fields contain less than 100 samples. This is likely to overfit the neural model, so it does not generalize well. Another reason could be due to the default word embedding model provided by Keras. In light of recent advances in pre-trained language models that rely on contextualized word embeddings [13], it is possible to fine-tune these models on a relatively small set of training data, which is a promising approach to beat the CRF model.

VI. Conclusion

We applied three sequence tagging models to extract metadata appearing on the cover pages of ETDs. Our best model (CRF-Visual) achieved 81.3%-96% F1 measure on seven metadata fields. Incorporating visual features into the CRF model boosts the F1 by 0.7% to 10.6% depending on metadata field. In the future, we will use pre-trained language models such as BERT [13] to initialize token representation learning. We will also add post-OCR error detection and correction into our pipeline when implementing the model on real data.

Acknowledgement

Support was provided by the Institute of Museum and Library Services through grant LG-37-19-0078-198.

References

[1] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp, "Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents," in *Proceedings of the 13th JCDL Conference*, 2013.

[2] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, "Automatic document metadata extraction using support vector machines," in *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '03, pp. 37–48, 2003.

[3] P. Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications," in *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'09, pp. 473–474, Springer-Verlag, 2009.

[4] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and L. Bolikowski, "CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature," *Int. J. Doc. Anal. Recognit.*, vol. 18, no. 4, p. 317–335, 2015.

[5] M. H. Choudhury, J. Wu, W. A. Ingram, and E. A. Fox, "A Heuristic Baseline Method for Metadata Extraction from Scanned Electronic Theses and Dissertations," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, p. 515–516, Association for Computing Machinery, 2020.

[6] I. Councill, C. L. Giles, and M.-Y. Kan, "ParsCit: an Open-source CRF Reference String Parsing Package," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[7] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, W. Peters, L. Derczynski, and et al., "Developing Language Processing Components with GATE Version 9 (a User Guide)," 2021. The University of Sheffield, Department of Computer Science, https://gate.ac.uk/sale/tao/split.html.

[8] J. Wu, S. R. Choudhury, A. Chiatti, C. Liang, and C. L. Giles, "HESDK: A Hybrid Approach to Extracting Scientific Domain Knowledge Entities," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1–4, 2017.

[9] S. D. Gollapalli and X. Li, "Keyphrase Extraction using Sequential Labeling," *CoRR*, vol. abs/1608.00329, 2016.

[10] J. Fonseca and K. Taghva, "Aligning Ground Truth Text with OCR Degraded Text," in *Intelligent Computing. CompCom 2019. Advances in Intelligent Systems and Computing, vol 997. Springer, Cham.*, pp. 815–833, Springer, Cham, 2019.

[11] M. Šošić and M. Šikić, ""Edlib: a C/C++ library for fast, exact sequence alignment using edit distance," *Bioinformatics*, vol. 33, no. 9, pp. 1394–1395, 2017.

[12] J. Wu, M. R. Ul Hoque, G. W. Reiske, M. C. Weigle, B. T. Bradshaw, H. D. Gaff, J. Li, and C. Kwan, "A Comparative Study of Sequence Tagging Methods for Domain Knowledge Entity Recognition in Biomedical Papers," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, JCDL '20, p. 397–400, 2020.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, vol. abs/1810.04805, 2018.