2021

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

Muntabir Hasan Choudhury
Old Dominion University

Himarsha R. Jayanetti
Old Dominion University

Jian Wu
Old Dominion University

William A. Ingram
Virginia Polytechnic Institute and State University

Edward A. Fox
Virginia Polytechnic Institute and State University

Original Publication Citation

Choudhury, M. H., Jayanetti, H. R., Wu, J., Ingram, W. A., & Fox, E. A. (2021). *Automatic metadata extraction incorporating visual features from scanned electronic theses and dissertations* [Paper presentation]. Proceedings of 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL '21). Virtual, Online.

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

Muntabir Hasan Choudhury
Himarsha R. Jayanetti
Jian Wu
Old Dominion University
Norfolk, VA
{mchou001,hjaya002,j1wu}@odu.edu

William A. Ingram
Edward A. Fox
Virginia Polytechnic Institute and State University
Blacksburg, VA
{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, so they often fail to extract metadata from scanned documents such as 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 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.

CCS CONCEPTS

• **Computing methodologies** → **Classification and regression trees**; • **Information systems** → **Digital libraries and archives**; • **Applied computing** → **Optical character recognition**.

KEYWORDS

Digital Libraries, Optical Character Recognition (OCR), Text Mining, Metadata Extraction, Heuristic Method, CRF, BiLSTM-CRF

ACM Reference Format:
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 *Proceedings of 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL'21).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3314111.3319916

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

JCDL'21, Sep 27–30, 2021, Virtual

ACM ISBN 9781450367097…$15.00
https://doi.org/10.1145/3314111.3319916

1 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 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 considered to extract metadata from scanned ETDs. Many tools [7–10] have been developed to automatically extract metadata for relatively short and born-digital documents, such as conference proceedings and journals 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. 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 incur a cost. Therefore, we adopted Tesseract-OCR, an open-source OCR tool, to extract metadata 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 our preliminary work, 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 a set of data with different a feature distribution. In this paper, we investigate the possibility of improving the generalizability of our method based on a learning-based model.

2 RELATED WORK

Many frameworks have been proposed to extract metadata from scholarly papers. CERMINE[10] 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. The model achieved an average F1 score of 77.5% for most metadata types and the benchmark evaluation outperformed the existing tools, including GROBID [9] and ParCit [2], 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 in the form of images will not be properly processed.

GROBID[9] is a text mining library for extracting bibliographic metadata from born-digital papers. GROBID is based on eleven different CRF models and each uses the same CRF-based framework which utilizes 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 but it fails to extract metadata from non-born digital documents such as the cover page of scanned ETDs.

In our previous work [1], 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. Table 1 shows the accuracy values obtained for each field for the sample of 100 ETDs. These range from 39% to 97%.

3 DATASET

The dataset used in our previous study [1] 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 and 150 non-STEM majors from 468 doctoral, 27 master's, and 5 bachelor's degrees. We derived the following seven intermediate datasets from our set of 500 ETDs.

(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 the Tesseract containing noisy text extracted from the TIFF images.
(4) TXT-clean: The cleansed version of TXT-OCR dataset after correcting misspellings, fixing the mistakes during OCR, 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 [3].
(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 all other universities from the ProQuest database.
(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).

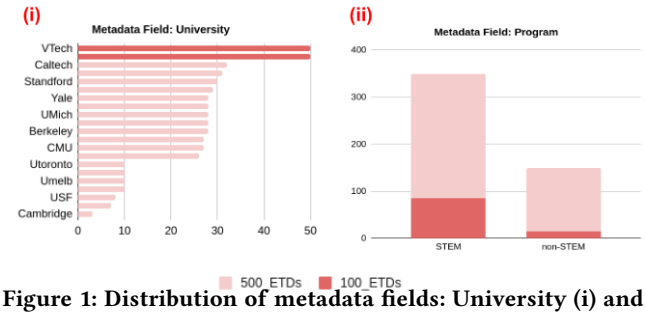


Figure 1: Distribution of metadata fields: University (i) and Program (ii) in the corpus of 500 ETDs.

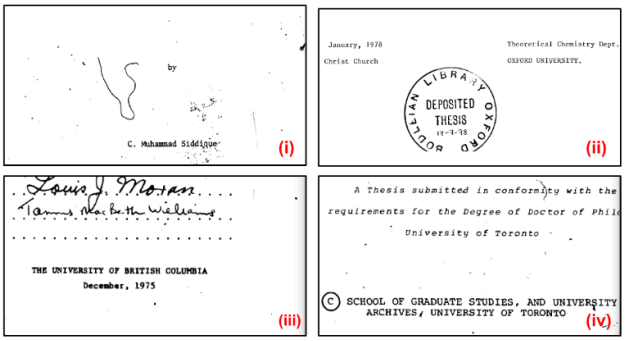


Figure 2: OCR challenges for ETDs: scribble (i), stamp (ii), overlapped letters (iii), and copyright character (iv).

(1) Some fields were not available on the cover page.
(2) Lines were present to fill the title, degree, author, etc.
(3) Multiple years are provided, such as "submitted year" and "publication year."
(4) There were ETD cover pages where author's previous educational certifications are listed (e.g., University of British Columbia) making it difficult to extract the degree field.
(5) College name is mentioned instead of university name (e.g., University of Oxford).

4 METHODOLOGY

4.1 CRF with Sequence Labeling (CRF Model)

CRF is a statistical modeling algorithm. This model assumes that the features are dependent on each other, but also considers future observation when modeling a sequence. It encodes each token of the annotated fields as the beginning (B) and inside (I). For example, if the token represents an author's name, we will tag it as B-author and I-author. The tokens which are not a part of the metadata fields should be tagged as outside (O). The BIO tagging schema has been applied in studies such as named entity recognition [11][2] and keyphrase extraction [6].

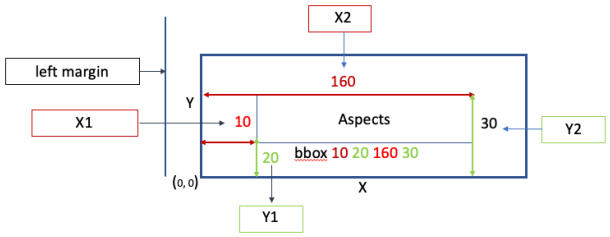
We tagged each token with Part of Speech (POS). POS tags are important here if the phrase consists of pronoun, preposition, or determiner, e.g., "at the Massachusetts Institute of Technology." If the current token is "Massachusetts" tagged as NNP, we can infer the POS of the two tokens before the current token. This assigns to the previous two tokens, "at" and "the," the POS tags IN and DT, respectively. Our model extracts the following features.

(1) Whether all the characters in the string are uppercase.
(2) Whether all the characters in the string are lowercase.
(3) Whether all the characters in a string 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 would have 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.

4.2 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 not only rely on text, but also 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 3(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 bottom margin to the bottom right corner of bbox, x2 is the distance from left margin to the upper right corner of bbox, and y2 is the distance from the bottom margin to the upper right corner of the bbox. All coordinates are measured with reference to the bottom-left corner of the token.



(a)

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

(b)

Figure 3: (a) Bounding box measurement (b) OCR output text (i.e., noisy) alignment with clean text

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 [5]. 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 [13] to align the 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 3(b) illustrates an example of the alignment. We incorporated three visual features to enhance the performance of the CRF model. The following three features have been normalized between 0 and 1.

(1) left margin — x1 as a feature for all tokens in the same line.
(2) upper left corner — y2 as a feature for all tokens.
(3) bottom right corner — y1 as a feature for all tokens.

4.3 BiLSTM-CRF Model

Bidirectional Long Short Term Memory (BiLSTM) can learn the hidden features automatically. It has been proven to be effective in sequence labeling problems [12] and in encoding sequence tokens into fixed-length vectors. BiLSTM tries to learn the context of the given sentence in both forward and backward directions. We added a CRF layer that classifies tokens based on their encoding. We investigated BiLSTM-CRF to extract seven metadata fields. The architecture of the classifier consists of three components: a word-embedding layer, a BiLSTM as an encoder, and a CRF layer. BiLSTM learns the forward and backward context in a sequence and feeds it into the CRF layer, which further predicts the labels for each token in a sequence. 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.

5 EVALUATION AND RESULTS

5.1 Heuristic Model

We apply the heuristic method on 500 variant sets of ETDs (Table 1). The accuracy of the heuristic method on the 500 ETDs is considerably lower than the accuracy in the 100 ETDs in the previous study [1]. This is because the new dataset contains ETDs from a more diverse set of universities and majors.

5.2 CRF Model and CRF-Visual

We 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 corresponding GT-meta and GT-rev. While comparing the title field against the GT-meta and GT-rev, a fraction of 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

Table 1: Performance (accuracy) comparison between the heuristic model on two datasets.

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%

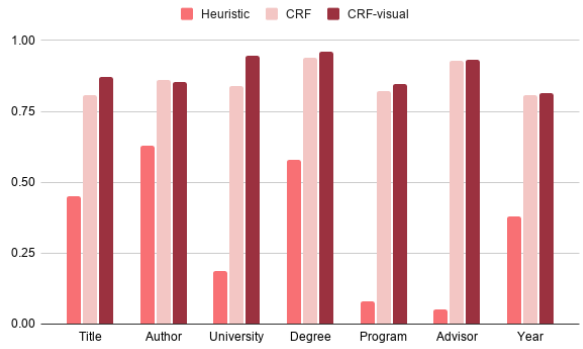


Figure 4: Performance (F1) comparison of the models

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 4 illustrates the performance of our model. The model outperformed the baseline model whereas CRF-visual outperformed both the baseline model and CRF.

5.3 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 4. 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 [4], 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.

6 CONCLUSION

We applied three models including CRF, CRF-Visual, and BiLSTM-CRF to extract seven metadata fields. We have observed that CRF-visual outperformed our Heuristic Baseline model and CRF with Sequence Labeling. Although BiLSTM-CRF did not perform well as we expected, we will fine-tune this model in the future by adding a pre-trained language models such as Bert. In the future, we will also add Post-OCR error correction to our model and run directly on real data (i.e., TXT-OCR) instead of rectified data (i.e., TXT-clean). Although we used our model on rectified data, it reflects the performance when the model will be used on real-world data.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Choudhury, M. H., Wu, J., Ingram, W. A., and Fox, E. A. 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* (2020), JCDL '20, Association for Computing Machinery, p. 515–516.

[2] Councill, I., Giles, C. L., and Kan, M.-Y. ParsCit: an open-source CRF reference string parsing package. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* (2008).

[3] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., Peters, W., Derczynski, L., 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.

[4] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018).

[5] Fonseca, J., and Taghva, K. Aligning ground truth text with OCR degraded text. In *Intelligent Computing. CompCom 2019. Advances in Intelligent Systems and Computing, vol 997. Springer, Cham.* 2019, pp. 815–833.

[6] Gollapalli, S. D., and Li, X. Keyphrase extraction using sequential labeling. *CoRR abs/1608.00329* (2016).

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

[8] Lipinski, M., Yao, K., Breitinger, C., Beel, J., and Gipp, B. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proceedings of the 13th JCDL Conference* (2013).

[9] Lopez, P. 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* (2009), ECDL'09, Springer-Verlag, pp. 473–474.

[10] Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., and Bolikowski, L. Cermine: Automatic extraction of structured metadata from scientific literature. *Int. J. Doc. Anal. Recognit. 18*, 4 (2015), 317–335.

[11] Wu, J., Choudhury, S. R., Chiatti, A., Liang, C., and Giles, C. L. Hesdk: A hybrid approach to extracting scientific domain knowledge entities. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2017), pp. 1–4.

[12] Wu, J., Ul Hoque, M. R., Reiske, G. W., Weigle, M. C., Bradshaw, B. T., Gaff, H. D., Li, J., and Kwan, C. 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* (2020), JCDL '20, Association for Computing Machinery, p. 397–400.

[13] Šošić, M., and Šikić, M. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics 33*, 9 (2017), 1394–1395.