Towards Aiding Research by Improving Access to Electronic Theses and Dissertations from Multiple Domains

CNI Fall 2021 Virtual Membership Meeting

Jian Wu (Old Dominion University Computer Science), William A. Ingram (Virginia Tech University Libraries), Edward A. Fox (Virginia Tech Computer Science)
Opening Books and the National Corpus of Graduate Research

• IMLS National Leadership Grants for Libraries
  • Investigating innovative ways machine learning and natural language processing can be applied to the national corpus of electronic theses and dissertations in order to extract knowledge, bibliographic and scientific data, and facilitate its identification, discovery, and reuse.

• Research Areas:
  1. Document analysis, information extraction
  2. Adding value through automatic classification and summarization
  3. User services – building better digital libraries
ETD Research Team at VT and ODU

Bill Ingram  
*Principal Investigator*  
Assistant Dean and Director of IT, University Libraries, Virginia Tech

Dr. Edward A. Fox  
*Co-PI*  
Professor, Computer Science, Virginia Tech

Dr. Jian Wu  
*Co-PI*  
Assistant Professor, Computer Science, Old Dominion University

Bipasha Banerjee  
*Graduate Assistant*  
Ph.D. Candidate, Computer Science, Virginia Tech

Muntabir Choudhury  
*Graduate Research Assistant*  
Ph.D. Student, Computer Science, Old Dominion University

And several notable students past and present (see some on next slide):

VT: Adheesh Sunil Juvekar, Eman Abdelrahman, Fatimah Alotaibi, Palakh Mignonne Jude, Sampanna Kahu, John Aromando, Gunnar Reiske
Other Notable Students

Sami Uddin
Graduate Research Assistant
Master’s student, Computer Science, Old Dominion University

Winston Shields
Graduate Research Assistant
Master’s student, Computer Science, Old Dominion University

Himarsha Jayanetti
Graduate Research Assistant
PhD student, Computer Science, Old Dominion University

Sampanna Kahu
Graduate Research Assistant
Master’s student, Computer Engineering, Virginia Tech

Palakh Mignonne Jude
Graduate Research Assistant
Master’s student, Computer Science, Virginia Tech

Eman Abdelrahman
Graduate Research Assistant
Master’s student, Computer Science, Virginia Tech
Outline

• Introduction
• Data Acquisition and Characteristics
• Repository, Searching, Services
• Language Model, Evaluations
• Results
• Conclusions, Future Work
Introduction

- Trend of increasing numbers of doctoral degrees in the United States
- Machine Learning (ML) and Deep Learning (DL): often data-driven
  - Pre-trained models in Computer Vision (CV), Natural Language Processing (NLP)
  - CV: VGG16, VGG19, ResNet-100, etc.: ImageNet (14 million), MS COCO (330k)
  - NLP: BERT, XLNet, etc.: BooksCorpus (800M), English Wikipedia (2500M), ClueWeb, Common Crawl
    - Self-supervised training: unlabeled data

ETD Collections

- NDLTD (Networked Digital Library of Theses and Dissertations):
  - 6+ million ETDs, 1.6 million in English
  - records only

- PDTG (ProQuest Dissertations & Theses Global):
  - 5 million ETD records, 2.7 million full text (majority in English)
  - subscription required

- Our collection:
  - ~450,000 ETD full text and metadata records

Numbers obtained on 5 November 2021
Building Our ETD Collection

• From over 40 university libraries; examples:

<table>
<thead>
<tr>
<th>Access type</th>
<th>University</th>
<th>URL Example</th>
<th>ETD set identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAI-PMH</td>
<td>Virginia Tech</td>
<td><a href="https://vtechworks.lib.vt.edu/oai/request?verb=Identify">https://vtechworks.lib.vt.edu/oai/request?verb=Identify</a></td>
<td>com_10919_5534</td>
</tr>
<tr>
<td>Sitemaps</td>
<td>UC system</td>
<td><a href="https://escholarship.org/siteMapIndex.xml">https://escholarship.org/siteMapIndex.xml</a></td>
<td></td>
</tr>
</tbody>
</table>

• Two access approaches
  • OAI-PMH -> metadata -> landing page -> PDF
  • Sitemaps -> landing page -> PDF

• Metadata
  • OAI-PMH
  • Scraping webpages, parsing HTML documents (obeying robots.txt)
Crawling Pipeline using OAI-PMH

OAI Portal

- ETD set identifier (e.g., com_10919_5534)
- Metadata prefix (e.g., oai_dc)

ETD list → XML → Landing page → Download

- Database
- File system

Crawl delay
Challenges and Lessons

- Not all PDFs are downloadable.
  - Restricted access
  - HTML DOM structure varies across repositories.
- Not all metadata has same fields of information.
  - Often missing: department, discipline, subjects, year issued
- Inconsistent formats, e.g., “date-issued”
  - 'mm-dd-yyyy'
  - 'before yyyy'
  - 'after yyyy'
  - 'yyyy strings'
  - missing
- Requests blocked even when obeying crawl-delay in robots.txt
## Our Top ETD Sources

<table>
<thead>
<tr>
<th>Top 10 Universities</th>
<th>PDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Ohio State University</td>
<td>55780</td>
</tr>
<tr>
<td>Virginia Tech</td>
<td>29597</td>
</tr>
<tr>
<td>Georgia Institute of Technology</td>
<td>22400</td>
</tr>
<tr>
<td>Texas State University</td>
<td>21702</td>
</tr>
<tr>
<td>Kansas State University</td>
<td>19299</td>
</tr>
<tr>
<td>The University of Texas at Austin</td>
<td>18283</td>
</tr>
<tr>
<td>Oklahoma State University</td>
<td>17746</td>
</tr>
<tr>
<td>North Carolina State University</td>
<td>15365</td>
</tr>
<tr>
<td>University of Illinois at Urbana-Champaign</td>
<td>14281</td>
</tr>
<tr>
<td>Rice University</td>
<td>13151</td>
</tr>
<tr>
<td><strong>Total (42)</strong></td>
<td><strong>451,358</strong></td>
</tr>
</tbody>
</table>
Distribution By Year

65955 ETDs dates not available in university-provided metadata

1875
ETD Repository

- ETDs have full text in PDF and XML.
- Total size: 3.4TB
- Hosted at ODU Computer Science data center
- Mirrored at VT University Libraries

Each ETD has a unique repository ID.
ETD Database

- Hosted in MySQL
- Born-digital vs. scanned
- Largest table: subject (1.9 million rows)
- Main table size: 451,358 rows

Database schema
<table>
<thead>
<tr>
<th>Field</th>
<th>Extracted Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>synthesis of well-defined single and multiphase polymers using various living polymerization methods</td>
</tr>
<tr>
<td>Author</td>
<td>joseph m. desimone</td>
</tr>
<tr>
<td>University</td>
<td>virginia polytechnic institute and state university</td>
</tr>
<tr>
<td>Degree</td>
<td>doctor of philosophy</td>
</tr>
<tr>
<td>Program</td>
<td>chemistry</td>
</tr>
<tr>
<td>Advisor</td>
<td>j. e. mcgrath h. w. gibson l. t. taylor t. c. ward g. l. wilkes</td>
</tr>
<tr>
<td>Year</td>
<td>1990</td>
</tr>
</tbody>
</table>

Choudhury et al. (2020 JCDL)
Choudhury et al. (2021 JCDL)
Metadata extraction from scanned ETDs

- ETD samples 1940-1990 (left figure).
- Conditional random field model with visual features (right figure)

Distribution of 500 scanned ETD samples over universities

Model Performance (81.3% - 97%) to extract 7 metadata fields.
ETD Search Engine

elasticsearch

MySQL

ETD repo

search

data

download

Web UI
Services with Collected ETDs

- Segmentation
- Summarization
- Subject classification
- Figure and table search (Kahu et al. 2021 JCDL)
- Research topic analysis (right diagram)

Courtesy: Ingram 2020 CNI presentation.

CS and Biology 2014-2018
Language Model

• LM power depends on training text.
  • BERT/RoBERTa/DistilBERT/XLNet/ELECTRA – Wikipedia, BooksCorpus, ClueWeb, CommonCrawl, Gigaword
  • SciBERT – PubMed
  • Specter – S2ORC (Semantic Scholar)
• Pre-trained LMs may not work well for ETDs on certain tasks.
  • The low vocabulary overlapping problem (e.g., discipline specific jargon) is seen in existing studies (e.g., Kandimalla et al. 2020)
• Training from scratch is expensive. → Fine-tune an existing LM
  • BERT (3.3 billion tokens, 52 hrs with a TPU; 7 days with a GPU)
  • SciBERT (3.2 billion tokens, 5 days with a TPU)
Building a Language Model for ETDs

- Start with SciBERT as the base model.
- Extract text from 8606 born-digital PDFs.
- 300 million tokens
- Fine-tune the LM using PyTorch transformers.
- Time taken ~20 hours on 1 Nvidia Quadro RTX 4000 GPU

- **Experiment 1**: 8606 documents (195 departments)
- **Experiment 2**: Remove the front matter (e.g., title page, table of contents, acknowledgements)
Evaluation

- Training with 8606 documents from 195 departments
- Measure: Perplexity
- Evaluation result: front matter confuses model -> excluded

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Training Text</th>
<th>Perplexity Score (the lower the better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language model 1</td>
<td>Full text</td>
<td>17.26</td>
</tr>
<tr>
<td>Language model 2</td>
<td>Full text but excluding the front matter</td>
<td>7.32</td>
</tr>
</tbody>
</table>

- Further evaluation regarding LM effect on classification
Conclusions

• OAI-PMH metadata gathering, crawling university libraries -> Collection of ~450k ETDs: full text PDF + metadata
• Analysis shows ETDs have inconsistent metadata information.
• Language model trained on 300 million tokens from ETDs
• Training uses fewer resources than general purpose (e.g., BERT) and scientific LMs (SciBERT) -> investigate whether it achieves comparable performance on subject classification
Ongoing and Future Work

- Improve metadata quality
  - Missing metadata (Choudhury et al. 2021 JCDL)
  - Fixing inconsistencies between metadata and data (table below)
  - Resolving university names: acronyms, full names, typos, e.g.,
    - Johns Hopkins University vs. jhu
    - Texas A&M University vs. Texas A & M University (additional space around “&”)

<table>
<thead>
<tr>
<th>ETD Cover Page</th>
<th>Library Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>chemistry</td>
<td>polymers</td>
</tr>
<tr>
<td>administration and supervision of special education</td>
<td>special education</td>
</tr>
<tr>
<td>educational administration</td>
<td>school administration</td>
</tr>
</tbody>
</table>
Ongoing and Future Work

• Improving LM
  • Increase number of training documents
  • Removing traces of tables and figures from text
  • Department names and disciplines in metadata -> unified schema (e.g., Microsoft Academic Graph) for subject classification

• Challenges
  • Conversion from PDF to text is not perfect.
  • Obtaining clean sentences
Ongoing and Future Work

• New features to be added to UI:
  • Multi-modality search results (chapters, figures, and tables)
  • Improved document summary page (right figure)
  • Improved services: metadata extraction, segmentation, classification, entity extraction, knowledge graph generation