Digitizing Economic History: Deep Learning Insights on Russia-U.S. Trade

Samuel Lin May 1, 2024

Abstract

This paper presents an end-to-end deep learning pipeline for digitizing and analyzing a historical dataset on U.S.-Russia commercial and trade relations between 1910 and 1963. The dataset, comprised of over 37,000 image files from the U.S. State Department Central Classified Files, documents significant geopolitical and economic interactions over a tumultuous period of history. This study seeks to transform these typewritten manuscripts from unstructured image files into an analyzable, structured digital format. Specifically, the deep learning pipeline employs the LayoutParser library [6] to extract document structures and the spaCy [4] library to categorize entities such as dates, currencies, countries, events, and locations. An evaluation of this preliminary work underscores the effectiveness of this pipeline and reveals directions for future work.

Introduction

Recent advancements in deep learning have revolutionized the ability to process large-scale unstructured data across various disciplines. However, this potential remains largely untapped in the study of economic history, primarily due to the unique challenges posed by the unstructured data. Although libraries and archives have scanned billions of pages of historical documents, many of these documents have unusual and complex structures that cannot be readily processed by rule-based and even deep learning approaches. As a result, these datasets are not widely used by researchers in economic history, even though the information in these historical records would benefit many of these researchers. Moreover, even if there are deep learning models that can effectively digitize these datasets, many researchers of economic history lack the technical expertise to build and employ these methods from scratch.

In this study, we focus on a dataset titled *Commercial and Trade Relations Between Tsarist Russia, the Soviet Union, and the U.S., 1910-1963* [8], obtained from the U.S. Department Central Classified Files. This collection of typewritten manuscripts varies greatly in structure and content. Generally, the documents relate to commercial and trade relations beginning in the Tsarist Russia period and extending through the Khrushchev period in Soviet history. Its contents contain a wide range of materials from U.S. diplomats including materials on treaties,

general conditions affecting trade, imports and exports, laws and regulations, customs administration, tariffs, and ports of entry activities. While this data is digitized, these records, which comprise over 37,000 image files, remain trapped as hard copy images, making it difficult to analyze at scale for research purposes. Each document in the source is linked to an Optical Character Recognition (OCR) conversion, but the conversions are very poor and have egregious errors. This is not surprising since the documents have a variety of complex layouts, which disrupt OCR and result in texts from different tables, rows, and text regions being incorrectly garbled together [5].

This study presents an end-to-end deep learning pipeline to convert this unstructured dataset into computable data for research analysis. All the code for this project, including the preliminary results on a subset of 500 images from 32 documents, is made publicly available [1]. First, the pipeline employs the LayoutParser library [6] to extract the various layout structures in the documents. Next, the pipeline uses the Google Cloud Vision OCR engine to convert the typewritten text into machine-encoded text. Finally, the spaCy library [4] is used to perform basic Named Entity Recognition (NER) on the extracted text. Our evaluation suggests that layout extraction can be improved by fine-tuning a model based on this specific dataset. It also underscores the effectiveness of layout extraction as a preprocessing step to prevent texts from different layout components from being incorrectly garbled together, thereby improving the accuracy of OCR conversion.

Background

This project was largely inspired by applications of deep learning techniques to unstructured economic data presented in Professor Melissa Dell's Economics 2355 course. There are several prior works that address related and/or similar problems. LayoutParser provides a comprehensive suite of functionalities to apply deep learning models to document image analysis tasks such as layout detection and character recognition. The pipeline presented by this project also employs LayoutParser [6] to streamline the extraction of document structures. A spectrum of pre-trained models is provided with LayoutParser, including the HJDataset [5], a robust dataset dedicated to historical Japanese documents with complex layouts. The HJDataset offers over 250,000 annotations across seven types of layout elements, addressing a lack of training data for documents in Asian languages. In this study, we use a model that was fine-tuned on the HJDataset to perform layout parsing. Although the HJDataset model is tailored for documents in Asian languages, it is readily accessible and serves as a useful performance benchmark for the U.S.-Russia trade dataset.

EffOCR [3] presents a novel open-source OCR package that approaches OCR as a character or word-level image retrieval problem, which significantly reduces the sample and computational resources needed for training and deploying OCR models. Future work might consider integrating EffOCR with this pipeline, although in this project, we choose Google Cloud Vision OCR for text recognition, leveraging simplicity in deployment as well as superior performance. Finally, this study performs basic NER on the extracted text. There is a large literature that relates on the analysis of text corpora. The famous *tf-idf* scheme measures the importance of a word to a document in a collection by assigning high value to words that have high frequency inside a document and are not found often in other documents. [7] Topic modeling is a type of unsupervised learning that seeks to discover the underlying structure in a corpus of text documents and categorize them into different groups based on their content. Latent Dirichlet Allocation [2], a notable technique in topic modeling, assumes that each document is a mixture of topics, and each topic is a distribution over words; moreover, it introduces a Dirichlet prior on the per-document topic and per-topic word distributions. In this study, we just perform simple NER on the extracted text using a pretrained model from the spaCy library [4].

Methods

The deep learning pipeline presented in this paper can be separated into two primary components: a *transcription phase* to convert the unstructured image files into a structured digital format, and an *analysis phase* to glean research insights from the structured text data in a systematic way. Prior to the transcription phase, there is also a *pre-processing phase* which seeks to improve the image quality of the dataset and thereby improve the resulting accuracy of the model.

Pre-processing Phase

In the pre-processing phase, the manuscripts are converted into JPEG format. Then each image is converted from the RGB color space to grayscale and binarized. This involves setting each pixel to either black or white based on a threshold value. This highlights the most prominent features in the images and prepares them for layout and text detection in the following phases. There is little loss of information from these pre-processing steps since all of the original documents are already in black and white. The images are also manually labeled with dates, which are provided by the original dataset source.

Transcription Phase

The LayoutParser library was used to streamline the entire transcription process. For layout detection, we choose a Detectron2 model that employs a mask R-CNN object detection al-

gorithm and was fine-tuned on the HJDataset. This model extracts seven layout components from the documents: page frames, rows, title regions, text regions, titles, subtitles, and other components. Next, the layout components are mapped to the corresponding image sections. OCR is then performed on each of these image snippets to extract the relevant text from each layout block. In our implementation, we use Google Cloud Vision's OCR engine.

Analysis Phase

In the analysis phase, NER is employed to identify and categorize key entities. We use the pretrained model `en_core_web_sm`, an English pipeline optimized for CPU, from the spaCy library. Specifically, this model identifies the following types of entities: organizations, geo-political entities, nationalities or religious or political groups, events, currencies, people, dates, locations, quantities, times, laws, and languages.

Results

Given the scope of this project and size of the dataset, it was difficult to perform a rigorous evaluation of the performance of this pipeline. We have limited the initial results to a small subset of the original dataset. This small sample contains 32 documents comprising 473 total image files, all of which are dated between 1950 and 1963. We hope that the following preliminary results will provide some insights into the performance of this deep learning pipeline.

A survey of the 473 images shows that the layout extraction misses many key components in the documents. There are several image files in which the layout extraction fails to identify any outstanding document structures, and others in which it misses entire headers or paragraphs. Nevertheless, among those components that are identified, most appear to be cohesive, e.g. paragraphs that are extracted indeed tend to be entire paragraphs. A count of the types of extracted components is shown in Table 1 in the Appendix.

Over 130,000 words were extracted from this small subset of the dataset. In general, the text recognition step seems to perform relatively well; despite occasional errors, the extracted texts are coherent, in contrast to the primitive OCR translations provided in the original data source. We observe that words may be missing letters or misspelled, but generally they are not garbled up as would be the case if layout extraction had not been performed beforehand. In general, the entity recognition seems to perform well, with entities being accurately categorized. The results as well as some examples of the extractions are shown in Table 2 in the Appendix.

The full results can be found in the project repository. [1]

Discussion

First, it must be acknowledged that further refinement of this preliminary pipeline requires a comprehensive, rigorous evaluation. To evaluate the layout extraction, the documents must be annotated with the correct components, so that the results of the layout extraction can be compared against a benchmark. To evaluate the text extraction, the documents must be manually labeled with the correct transcriptions. Similarly, the entity recognition step requires that each text component be manually tagged with the correct components.

The shortcomings of this initial pipeline reveal several important areas for future work in this project. Specifically, the layout recognition step often fails to recognize some document structures. I believe that it is worthwhile to fine-tune a model specifically for this dataset, just as a Detectron2 model was finetuned on the HJDataset. Of course, this would require manual annotation of the dataset, which can be labor-intensive. Additionally, the integration of more sophisticated entity recognition tools could enhance the depth of data analysis and allow for more intricate insights into the diplomatic language and trade terminology used during this era.

Despite the various limitations, this initial pipeline already yields promising results. For example, an initial survey of the extracted entities supports our understanding of Cold War geopolitics. It would be interesting to perform fine-grained analysis of these entities. For example, a map of the geopolitical entities might provide insights into political alliances and economic policies that shaped U.S.-Soviet trade relations. An analysis of the currencies mentioned might help researchers understand the economic dynamics and financial strategies employed by these nations during the Cold War era.

In conclusion, the development of a deep learning pipeline for digitizing and analyzing historical U.S.-Russia trade documents presents a significant advancement in economic history research. By transforming these documents into a structured digital format, this project enables a more detailed and efficient examination of historical interactions. Despite some limitations in accuracy and the need for further refinements, the initial results are promising and suggest a substantial potential for deep learning technologies to uncover new insights from archival data. Continued improvements in layout recognition, text extraction, and entity analysis, along with the expansion of the dataset to include a wider range of documents, are expected to improve the pipeline's performance. This project supports the work of historians and researchers to understand complex historical trade relations.

Appendix

Table 1: Components Extracted from Documents

Component	Count
Page Frames	358
Rows	823
Title Regions	50
Text Regions	3
Titles	9
Subtitles	4
Others	133

Table 2: Entities from Documents Identified and Categorized

Entity	Count	Examples from Extraction
Organization	2625	'The Tunisian Government', 'the Munitions Board'
Geopolitical Entity	2861	'Singapore', 'United Kingdom'
Nationality/Religious/Political Groups	1864	'Soviet', 'French', 'American'
Events	46	'World War II', 'the Algerian Revolution'
Currencies	266	'francs', 'dinars'
People	1159	'Heinrich Rau', 'Perron', 'Dashkevitch'
Dates	1575	'July 15, 1954'
Locations	322	'North Vancouver', 'West', 'Siberia'
Quantities	343	'3,000 metric tons'
Times	43	'11 a.m.'
Laws	33	'the Trade Agreement Act', 'the Neutrality Act'
Languages	22	'English', 'Russian', 'Spanish'

References

[1] https://github.com/samuellin01/ec2355-final-project.

[2] David M Blei, Andrew Y Ng and Michael I Jordan. 'Latent dirichlet allocation'. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[3] Tom Bryan et al. 'EfficientOCR: An Extensible, Open-Source Package for Efficiently Digitizing World Knowledge'. In: *arXiv preprint arXiv:2310.10050* (2023).

[4] Matthew Honnibal et al. 'spaCy: Industrial-strength Natural Language Processing in Python'. In: (2020). DOI: 10.5281/zenodo.1212303.

[5] Zejiang Shen, Kaixuan Zhang and Melissa Dell. 'A large dataset of historical Japanese documents with complex layouts'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 548–549.

[6] Zejiang Shen et al. 'Layoutparser: A unified toolkit for deep learning based document image analysis'. In: *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*. Springer. 2021, pp. 131–146.

[7] Karen Sparck Jones. 'A statistical interpretation of term specificity and its application in retrieval'. In: *Document Retrieval Systems*. GBR: Taylor Graham Publishing, 1988, pp. 132–142. ISBN: 0947568212.

[8] U.S. State Department. *Commercial and Trade Relations between Tsarist Russia, the Soviet Union and the U.S., 1910-1963*. Online Archive. Central Classified Files, Records of the U.S. State Department, National Archives, College Park, MD. Decimal numbers 661.11 and 611.61 for the period 1910-1949, and 461.11 and 411.61 for 1950-1963. 1910–1963. URL: http://gdc.gale.com/archivesunbound/.