University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Library Conference Presentations and Speeches

Libraries at University of Nebraska-Lincoln

9-26-2018

Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures

Elizabeth Lorang
University of Nebraska - Lincoln, llorang2@unl.edu

Leen-Kiat Soh
University of Nebraska-Lincoln, lsoh2@unl.edu

Lorang, Elizabeth and Soh, Leen-Kiat, "Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures" (2018). *Library Conference Presentations and Speeches*. 143.
http://digitalcommons.unl.edu/library_talks/143

Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, Chulwoo Pack, Delaram Rahimi

National Digital Newspaper Program Meeting

September 26, 2018

Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures

Introducing the Team and Our Work

Image Analysis for Archival Discovery, or Aida, is a cross-disciplinary research team, with researchers from computer science, literary studies, and library and information science. Our work explores the question,

> What might we learn about digital collections of cultural heritage materials, and how might we augment use and access of these collections, if we focus on the digital images being created as librarians, archivists, museum professionals and others are digitizing cultural heritage materials?

We are particularly interested in this question in relation to *textual materials*, such as records and manuscripts and especially heterogeneous collections and materials, such as periodicals and newspapers.

We believe that researchers and developers might do far more with digital images at all stages of digitization and use, and that attention to the digital images will yield greater understanding across a range of domains. There is much to learn about the materials themselves as well as about processes, including the values we bring to digitization and those that get enacted through digitization, as well as how current work builds on those values previously enacted in keeping and maintaining the original items themselves. Studying these processes can function as a form of social and cultural history, of interrogating how society and culture are represented in collecting and in technology. We see exploration and study of digital images of historic materials as a mode of asking questions about the materials across their many forms, from physical originals, to microform duplications, to digital copies, represented as images, text, and metadata.

In our presentation today, we will introduce you to the work-to-date of the Aida team, including analyses we are in the midst of running. We'll start with a broad question and begin to explore that question through a much more specific challenge, and from there zoom back out to the from the challenge to the larger landscape. The goal is to situate our work broadly as well as hinge it on some very specific challenges and in doing so demonstrate the many types of questions and domains to be explored in digitized newspapers. We have tried to pitch this presentation so that it will be intelligible and interesting to a broad audience interested in historic newspapers and digitization. For today, we are intentionally leaving out the more detailed look at specific algorithms and step-by-step processes, but we can talk about those things if you're interested during the Q&A. For the prepared remarks, though we have abstracted the specifics to their foundational concepts and ideas.

The question, "What might we learn about digital collections of cultural heritage materials, and how might we augment use and access of these collections, if we focus on digital images," is a big one. We therefore carved out a smaller, but still ambitious, question that would begin to address the larger one: Is it possible to use visual signals in historic newspapers to identify content by form or genre? Specifically, can we find poetic content in historic newspapers based on visual signals alone?

In the discussion portion of our time today, we'd be happy to talk more about *why poetry*. For now, we will in brief that we focus on poetry, or poetic content, for several reasons: scale, visual distinctness, and interest and significance. There are **millions** of poems in historic newspapers, they are visually distinctive in the newspaper pages, and there is growing interest in newspaper poems across various subdomains in literary studies, including book history, women's writing, and cultural studies.

Our principle strategy to date has been to employ image processing techniques to generate data about visual features from the newspaper pages and then use those extracted features within a computational system called an artificial neural network. The underlying idea is to see if we can train, and then test, a machine learning system to recognize poetic content based on the features we extracted from the digital images, and then to extend such a system to identify other types of content in newspapers based on visual features. We achieved promising results with this technique in 2015, as we wrote about at the time in *D-Lib Magazine*. In the time since, we have focused on improving the efficacy and also the scalability of the system.

Currently, we have a National Digital Platform research grant from the Institute of Museum and Library Services, and the goals of that grant are to: extend Aida's software across a more diverse range of digitized newspapers and textual forms and assess the broader potential of image analysis as a methodology for information classification, identification, discovery, and retrieval in digital libraries. Specifically, we are working to 1) analyze and verify our image analysis approach and extend it so that it is newspaper agnostic, type agnostic, and language agnostic; 2) scale and revise the intelligent image analysis approach and determine the ideal balance between precision and recall for this work; 3) distribute metadata and develop a new digital collection using the extracted content; and 4) disseminate results, including adding to the scholarly literature on these topics and providing training for members of library and archive communities.

Ongoing Efforts to Improve Our Approach

Dealing with Noise

Our largest case study has been to test the effectiveness of our first-generation approach on page images from Chronicling America from 1836-1840, an image set that we downloaded in 2016. When we tested the system that we used for our *D-Lib* article on more than 8,000 pages from this time period, the results lagged well behind what we saw in our smaller proof-of-concept project. Why? As we looked at the results and at the various files created throughout our process—newspaper page images, segmented newspaper page images, post-processed segments, files with feature information measured from the snippets—we determined several points in the process that were potential areas of weakness. We found that our existing approach was highly susceptible to noise in the digital images. This noise was caused by features such as bleed-through, low contrast in the images, and occluding "blobs," which prevent our system from being able to accurately capture the visual cues. One lesson, therefore, is that if we want to be able to use the digital images for the types of purposes we imagine, we need to be able to handle the "noise." In response, we have been developing approaches that minimize the impact of this noise for identifying and measuring visual features.

One strategy would be to remove entirely the noise from the images of the newspaper pages. However, the degrees and amounts of noise on the pages are significantly varied – that is the signal-to-noise (SNR) ratios in these images are not consistent due to the differences in qualities of the physical newspaper pages, in the microphotographic reproductions, and in the digital scanning. A particularly aggressive noise removal or reduction technique could easily damage the non-noise pixels causing inaccurate feature extraction. Our goal, then, is not to eliminate the noise, but rather to improve our binarization of the image into "object" and "background" information (represented as black and white pixels, respectively) and extract visual structures with a noise-tolerant text line segmentation approach.

Improving Performance: Binarization

We compared this new approach with our first-generation approach on a set of 17,000 image snippets from the 1836-1840 case study. The highest testing accuracy of our initial approach was 61.89% (that is, just 61.89% of image snippets were classified correctly as containing or not containing poetic content), while we achieved 73.1% with the improved way of handling noise in the images. We also improved our recall rate from an average of 62% in our initial approach to an average of 76.4% in our improved approach. That means that we successfully returned 76.4% of the poetic content in the snippets in our improved approach, as compared with just 62% of the poetic content in the initial approach.

Improving Performance: Deep Learning

We are working on alternative approaches that will further improve our accuracy in identifying poetic content: one approach we are exploring has yielded initial results of accuracy achieved 92.4%. We also observed a significant improvement on precision to 92.8% and recall to 91.9%, so we were recalling just under 92% of the poetic content in the snippets. The initial Convolutional Neural Network model we explored learned to recognize a snippet with poetic content in about 50 epochs. Considering there are only 50 snippets per epoch, the Convolutional Neural Network model learned the dataset of 17,000 snippets using only 15% of the snippets. This result is not entirely surprising since Convolutional Neural Networks are known to improve image recognition significantly. Then, for the future work, we will explore deeper networks to seek even higher classification accuracy. At the same time, due to the initial Convolutional Neural Nets test suggesting that it may be able to learn with a smaller training set, we are also interested in pushing the training set even smaller than 15%.

Improving Performance: Page Segmentation or Zoning

Improvements on feature extraction and classification, though, only affect the newspaper page images that advance to that point in our process—those that get to the feature extraction and classification stage of our end-to-end process. You'll recall from a few minutes ago that the first part of our process takes in a newspaper page and cuts it into dozens of sub-images from the whole newspaper page. Our first-generation approach to generating page image "snippets" has been to feed in an image of a newspaper page, find the newspaper columns present on the page, and then cut each column into a series of column snippets of a fixed width:height ratio. Our premise was that we could ultimately take the snippet, determine whether it featured poetic content, and then, if so, connect that information with coordinates from the page image, to determine more locally where on the page the poetic content appeared. We were drawn to this approach because it was computationally cheaper and faster than approaches that employed more extensive processing—for example geometric analysis or logical layout analysis.

In the best of cases, this approach yields very good image snippets for use in the rest of our process. We noticed early on, however, that a variety of factors influence our ability to create good image snippets, and bad-quality snippets adversely affect our ability to train and deploy a classifier for identifying content. We consider snippets to be of good quality if they contain text/content from only a single column of a newspaper and include the full-width of that column's content. Bad snippets include content from multiple columns, including multiple incomplete columns or a complete column with content also from a secondary column.

To decrease the number of bad snippets that would make it through to feature extraction and classification, we created a set of tests that we use to evaluate a page before cutting it up into snippets. One test checks to see whether we can determine column breaks, by the presence of either continuous black or white lines marking columns. We initially also included a test to check whether the page contained more than two columns. If columns are found, we then evaluate whether the columns are of roughly the same width across the page. Page images that pass all three of those tests then move on to the actual segmentation stage.

In our first implementation of the rules, just under 41% of page images passed the tests and moved on to segmentation. In the second implementation of the rules—where we cut the test that checks to see if there are more than two columns—the number of page images that passed the tests increased to just over 45%, with those page images then moving on to segmentation. This means in practice, of course, that well over 50% of page images get discarded before we even begin looking for poetic content. Why?

In addition to the statistics we record for each image as it goes through the pre-segmentation tests—statistics about the number of columns, deviation in columns, and so on—we also conducted an observational analysis of the images to see why our algorithms may have been having a difficult time finding column breaks and columns of a column width. We observed that images that failed on average have lower contrast, the newspaper pages have more graphical content, more text spans the page, and there is more "noise" from what we have been calling (in a very technical term) "blobs."

As a result of this work and preliminary analysis, we have been revising our approach to segmentation to implement a more robust approach to zoning that may be less susceptible to these issues. We are drawing on layout analysis and zoning in this revised approach. OCR systems perform layout analysis and zoning as a pre-processing step for text recognition. It is built in to OCR web services such as ABBYY Cloud and in the open source Tesseract OCR engine, for example.

Some of this zoning information is maintained in the OCR XML files made available through Chronicling America, including zoning coordinates for strings, text lines, some larger text blocks, and page-level information. Layout analysis and zoning remaining challenging for historic newspapers and periodicals, particularly those with dense and mixed layouts. This topic—newspaper segmentation—has even been the focus of a recurring contest in the document analysis community because of the challenges it poses. Off-the-shelf products have increased their effectiveness in dealing with newspaper layout, but challenges remain.

Our current strategy, then, is focused on expanding regions around text-lines to identify textual zones. We also know that given our project's aims, we can compromise to an extent between the accuracy of zoning and the processing time required. Our aim is to find a zoning approach that creates output we can work with for visual analysis, and it does not need to exceed that specification. One outcome of our work, then, may be a rubric of sorts for the zoning most useful for supporting different types of study and analysis with digitized historic newspapers, not only zoning for OCR. For our work, for example, we know we do not need 100% perfectly accurate zoning. Perfect zoning would mean that a zone contains only the semantic content of its article and nothing more or less. While we do not need perfection, we do need to be able to accurately segment a newspaper page into zones that (1) do not prematurely end the horizontal visual field; and (2) do not zone images into zones so small that visual features cannot be extracted.

In addition to extending this work, a next step will also be extracting information about and analyzing the text blocks declared in the OCR XML files and segmenting the image pages based on those zones (i.e., using the bounding box coordinate information to cut a page into zones and extract those zones from the page). We can then explore the application of multiple of our feature extraction and machine learning approaches to the segments created via different means.

Studying Newspaper Pages in Chronicling America

What does it mean for an approach to be language agnostic?

In the rest of our time today, we want to turn to some other analyses we have been recently working on. To introduce this work, we'll start with a brief story: In April, I participated in a symposium at the American Antiquarian Society on multi-ethnic periodicals. In preparation for that event, I started looking a bit more at how our approaches to date worked on our dataset as a whole as well as how they worked when examined by the language of the newspaper and from particular ethnic and racial groups that were creating the newspapers or who were primary audiences for the newspapers. Up until that point, we had been taking our 1836-1840 dataset as a whole and had not broken it down by various criteria, except to look at the visual features of pages that seemed to affect our work. When we began to look at the results with regard to language of the papers, the results were surprising: Significantly, most pages in languages other than English never made it passed the segmentation tests. Pages that don't pass the segmentation tests do not go the rest of the way in the processing, meaning we would not have found poetry in languages other than English, even though it is present in the dataset.

Breaking this down a bit further: When we evaluated the 21,000 page images from 1836-1840 according to our initial segmentation rules, we knew that a total of 40.8% of the pages passed the pre-segmentation tests and proceeded through the rest of the segmentation, feature extraction, and classification processes. What happened when we looked below this statistic as a whole and looked at the stats by language of the newspaper pages?

Pages from newspapers that were in English passed at a slightly higher rate than the total at 43.7%, with pages in other languages passing at only about 25%, when assessed together. Language-by-language, there are stark differences by language: While over 79% of pages in German passed the pre-segmentation tests, only 11.6% and 10.5% of pages in Spanish and French passed the tests, respectively.

You may recall from earlier in the presentation that we altered our pre-segmentation tests to account for two-column newspapers. When we again evaluated the 21,000 page images from 1836-1840 with this change, we improved segmentation across the complete set to 45.5% of pages passing the pre-segmentation tests.

Pages in English again passed at a slightly higher rate of 47.1%. Pages in languages other than English increased their rate to 36%, with German pages at 79.8%, Spanish at 26.8%, and French at 12.4%. Taken together, however, our pre-segmentation tests still eliminated the significant majority of pages in languages other than English: pages that would not continue on to further processing and would not be evaluated for the presence of poetry.

To be clear, it is not the actual language of the pages themselves that is affecting the results: at no point are we evaluating the language of the papers or do anything with the linguistic content—all of our measures for segmentation and other analysis are of pixels. However, something about the pages appears to be disproportionately affecting segmentation results for pages in languages other than English—whether positively, as in the case of the German pages, or negatively, as in the case of the French and Spanish. If we imagine a future where people might use tools such as ours to generate new corpora for study out of collections such as Chronicling America, therefore, with our current system we might inadvertently find ourselves in a future where those corpora over-represent some languages and under-represent others, even when one of the great promises of digitization and access is the work of recovery and reclamation.

When we proposed our project first to NEH and then to IMLS, we highlighted the fact that our approach should be, for all intents and purposes, language agnostic: As long as the newspaper content was visually similar across the newspapers in Chronicling America and other corpora, we should be just as able to identify poetic and other content across all languages represented. As part of our current research grant, we need to expressly test this idea.

In a very literal sense, our approach is language agnostic: we never actually evaluate the characters as linguistic characters or attempt to process the language in any way. What would it really mean, however, for our approach to be language-agnostic? This question is one that has grown in complexity for us once we started looking at how well our approaches were working for pages in specific languages. To be functionally language agnostic isn't about the language at all but attention to the range of factors that affected the materiality of the papers, their production, keeping, filming, digitization, and the way the histories of the materials are present and encoded at these different stages.

Conditions of Creation & Legacies of Care

Seeing this peculiarity in our segmentation data from 1836-1840, we determined to explore this further over a large set of languages and also ultimately at a larger scale, with the idea that as part of the work we needed to study the images themselves more holistically—not only process them to work for our purposes. We determined that we needed to do higher level and corpus wide analysis of Chronicling America for different features.

Our questions include: What features are more or less present in these pages that are affecting our ability to segment them? Do those features in any way group by the newspapers' original language or by the newspapers' racial or ethnic communities, as they appeared to do in our initial test? If so, what does this data tell us about the newspapers themselves or how the newspapers have been collected, cared for, copied, and digitized? Are we seeing the impact here of legacies of care on our processes and methods? Is this a moment to, as Bergis Jules has framed it, "[confront] our failure of care around the legacies of marginalized people in the archives"? Are we seeing the visible signs of how these newspapers were *not* cared for over time, in their original form and in duplicative processes? While digitization is often positioned as a corrective, it also amplifies these legacies of prior treatment, and machine learning systems amplify biases and feed those biases into systems for future learning.

Even as we will continue our work on image zoning and segmentation and classifying images with neural networks, we are also taking now a higher-level view to analyze various newspaper corpora according to a range of features. While the full range and scope of this work is evolving, we have started with a subset of images from Chronicling America to study and test our approaches. We began by working with the Chronicling America team to get a copy of the Chronicling America database so that we could generate page-level statistics for the corpus (for example, how many pages are in the collection for each year, by geographic location (city and state), and by language? For our purposes, we needed this information at the page-level, not only at the issue-level, for example. While we ultimately want to explore a host of visual features in relation to language, ethnicity, time, and geography of the papers, we are starting with language, because of what we have already seen in relation to language and segmentation.

Results and Analysis of the Page-Language Case Study

Because of the connection to language in our earlier segmentation test, we are focusing on analysis by language for this first round of analysis. For the first set of tests, we decided to evaluate a sub-corpus with an equal representation of newspaper pages in English and newspaper pages in languages other than English. To determine the subset, we looked at the distribution of pages by language in Chronicling America over time. We focused on newspapers that were identified with a single language only, so that when we conducted our sampling, we would get newspaper pages in the language we expected.

With these criteria in mind, we saw that single-language newspaper pages fall between 1834 and 1922, inclusive, in the current Chronicling America corpus (as of August 16, 2018). We began with a test set of 20 pages from each year, 10 in English and 10 in languages other than English, with a roughly equal distribution of pages across all of the languages represented in that year in Chronicling America. This meant that our dataset was balanced with regard to pages from English-language papers and from non-English-language papers, but it is not balanced in terms of the number of pages across all languages represented (something we would want to do in the future).

We first ran this new set of images through our pre-tests for segmentation. For now, we see these segmentation pre-tests as a proxy for some measure of complexity or "non-standard"-ness of the images. Proxies are tricky of course, but for now the segmentation tests give us a first sense of where we might dig deeper.

Segmentation pre-test results here shared some similarities but also highlighted some differences with the 1836-1840 set. For some languages, no pages passed the segmentation tests. These languages (Hawaiian, Lithuanian, Slovenian) also had some of the smaller number of pages in this set. Pages from German-language newspapers again passed the tests at a percentage significantly higher than the rest of the pages—though at a much-reduced percentage than from the 1836-1840 test, where they were passing at nearly 80%. Conversely, we saw a marked increase in the number of pages from French-language papers that passed the segmentation tests, while the number of pages in Spanish-language papers dropped by a few percentage points. There is significant variation in the percentage of pages that pass segmentation when broken down by the language of the pages.

However, we also observe significant variation in the percentage of pages that pass segmentation when we consider the pages over time as well as over their sequence in their larger newspaper issue.

We know from our earlier work that contrast, skew, range effect, bleed-through and "blobs," and multi-column content affect our work at various stages. As a result, we started with some of these features as features to analyze—as features that might tell us something about the pages and the corpus.

As a starting point, we have calculated and analyzed each page and grouped results by language for several features: contrast, range effect, orientation skew, "noisiness," and complexity of the page.

When we evaluated the set for contrast and range effect, we see that most newspapers have contrast values between 40 and 80, with contrast values above 30 demonstrating good contrast. The lower the contrast value, the worse the contrast. In the current test set, only between 6-7% of the images have bad contrast, according to our measures.

The ideal value for range effect is 0, meaning that no range effect is present. Range effect becomes challenging when its value is greater than 3. With our current test set, just over 25% of the images in the set have range effect values greater than 3, meaning that range effect is potentially problematic in ¼ of the images in the current set.

We also observe across the set that contrast is pretty consistent regardless of the newspaper page's language. Contrast also stays pretty consistent over time. Range effect, on the other hand, not only varies across pages in different languages, it also changes over time for each language.

Histograms of the data also illustrate this: The data on Contrast are pretty symmetric, while the Range effect data are right-skewed. We can also see that there are a lot of outliers in our Range effect data. The presence of outliers in our range effect data indicate that the mean (average) would not provide a good estimate of what language has the highest/lowest Range effect. Instead, median would give us a good estimate to evaluate the center of the data.

We anticipated that skew might have a significant impact on our ability to segment pages and also in later classification processes. As an initial test, we measured the global orientation skew of the pages in the set. Orientation skew ranged from -2 (counter-clockwise skew) to +2 (clockwise skew). Because our measure of global skew considers the background on which a newspaper page was microfilmed, there is little orientation skew overall. A more effective measure is likely to be local skew, relative to particular parts of the page, or other measures of warpedness or beveled nature of the page. Part of our evolving work on zoning deals with local skew, so in the near term, we will explore local skew at scale—not only as something that we account for in zoning.

To this point in our analysis, then, the most problematic feature of the pages for a range of processes appears to be the presence of range effect. From our earlier work, and our work on zoning, we also know that bleed-through from reverse of a newspaper page and a non-textual "blobs"—typically damage or staining of some sort—posed significant challenge. We set out, then, to explore these features as what we called the "noisiness" of the page.

Assessing effects of bleed through, blobs (e.g., stains), and other non-textual artifacts

- Defects or degradations of a page, or of the digitization process

Based on histogram analysis—of pixels' intensity values—of each page

Our current technique:

- If an image is deemed noisy, then it is indeed noisy
- If an image is deemed NOT noisy or of low noise then it is either (1) indeed NOT noisy or of low noise, or (2) of poor contrast

Finally, we compute the number of characters, lines, and zones using zoning technique, as a way to begin measuring both complexity and density of the textual content on the page.

Numbers of characters, lines, and zones vary across newspapers of different languages

- Polish, Italian, Czech: more characters per page than Spanish
- Italian, Polish: more zones per page than Spanish, Czech

Polish-language newspapers had more outliers; Italian-language newspapers more consistent

Italian- and Czech-language newspapers had more characters per line

- Does it mean they had more compact typesetting?

- To investigate further: average length of lines

Spanish-language newspapers had more lines per zone

- Does it mean Spanish-language newspapers had more uniform layouts such that a detected zone was not broken up into multiple smaller zones?

To investigate further: average size of zones

Next Steps & Take-Aways

Next Steps

- Extend the page-level exploration to > 30,000 pages and analyze according to language, time, and geography
- Extend page-level exploration to Burney Collection and other corpora
- Continue exploring zoning options, including region-growing approach and using zones identified in CA XML
- Perform larger test of deep-learning classification approach for classifying textual content, analyze, and refines

Take-Aways

- The nature and type of research that can be done with digitized newspaper collections at the size of Chronicling America and NDNP goes far beyond finding information in the pages and extends to many disciplines
- Digitization encodes values and beliefs and can amplify traces of previous treatment. One question that remains open for us, and one we continue to study is which features that affect our work are inherent to the newspapers themselves, and which are the result of time and intervening processes. Our work thus far has shown the importance of features

introduced during microfilming and digitization, such as contrast and noise for image-based approaches.

- There may be no one-size fits all approach, even when dealing with a single type of material such as newspapers
- We have an obligation to critique processes and see what, and who, they are leaving out. The results of those processes will shape what people study and the questions they pursue.
- No shortage to the questions to be pursued, or the domains of knowledge that can be informed by, through the work and products of the NDNP and Chronicling America.