Allen, R.B., Waldstein, I., and Zhu, W., Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres. ICADL, 2008, 380-387

Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres

Robert B. Allen, Ilya Waldstein, and Weizhong Zhu

College of Information Science and Technology, Drexel University
rba@drexel.edu, imw22@drexel.edu, wz32@drexel.edu

Abstract. Many historical newspapers are being digitized. We aim to support access to them via text analysis of the OCRd content. However, the OCR includes many errors; so extracting meaningful content from it is difficult. A pipeline of processing steps is proposed. Here, we describe the first two steps: segmentation and genre identification. The segmentation procedure based on headings was quite successful. Genre identification worked well for easily defined genre categories such as weather reports. We also propose additional techniques which may improve the accuracy still farther.

Keywords: Categorization, Genres, Historical Newspapers, OCR, Segmentation.

1 Introduction

1.1 Digitized Collections of Historical Newspapers

Historical newspapers are an important source of past local news. They are being digitized to preserve content and enable search and retrieval. In the U.S., the National Digital Newspaper Project (NDNP) has several hundred thousand pages of historical newspapers digitized from microfilm, processed with OCR, placed online, and made accessible through search-engine interfaces. These materials should be more browseable if individual components are identified. Because of the large volume of material this would be costly to do manually, however, automated processing could be used to augment or replace manual classification. While there has been other work on processing modern news articles, there has been surprisingly little text processing research on historical newspapers or even on newspaper in their entirety. Because of the difficulty with OCR, our goal is not perfect recognition but enough improvement in accuracy that the process will eventually converge. We do not aim to develop new algorithms but to find robust solutions to a practical problem.

1.2 Automated Processing with the Pipeline

There are four main steps in the classification process (Figure 1). The modules take into account the properties of the data set, such as the date and page number, and the results of the previous processing step. We have developed an ad hoc XML wrapper for each of the newspaper segments which identifies the date, page, and coordinates of each

G. Buchanan, M. Masoodian, S.J. Cunningham (Eds.): ICADL 2008, LNCS 5362, pp. 380–387, 2008.

segment. The wrapper accumulates information as it passes down the pipeline. For example, once the genre is assigned it is stored in the XML wrapper. Eventually, this ad hoc XML scheme could be formalized as an article-level METS description.

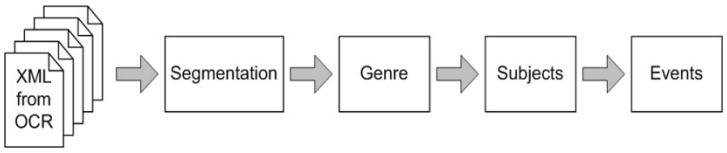

Fig. 1. Pipeline for processing the OCR'd text

The pipeline provides a convenient model for conceptualizing these complex processes. However, it is a simplification in regard to the distinction between genres and subjects. The vagueness between genres and subjects makes a clear separation between them difficult. For example, a description of a baseball game can fall into the genre of sports notices or it can be counted as an article about baseball.

1.3 Test Corpus

NDNP files for the years 1900-1910 from Washington DC were obtained. From those, we focused on the *Washington Times* for 1904. Because they conformed to the NDNP specification, they included OCR which was wrapped into METS Alto files [7]. Thus, in addition to the OCR text itself, other attributes were included such as the coordinates for each word and the fonts. It is worth noting that the Alto files differ across digitization projects in the level of detail they include; for example, a few digitization projects have developed Alto files that include keystroked article headings and image captions. As with many newspapers, there is substantial variation from year to year. The OCR for 1904 was selected as a moderately difficult data set due to its quality. Figure 2 shows text samples from two OCR records from our test collection. To reduce the introduction of errors, only minimal processing was applied to the original OCR files. Thus, some corrections are relatively easy to make while others are more difficult. Indeed many other researchers have posited that OCR of this quality is too complex to process automatically. We argue that there are sufficient constraints that allow this text to be processed automatically. In particular, we believe that the tasks which involve categorization such as genre and subject identification will succeed even though some of the words in the OCR are not intelligible.

STATEHOOD MEASURE WILL PASS THE HOUSE Republicans Determine to Rush Hamilton Bill Through to Be Ready for Senate in December margin but NSW Mexico which hasnearly I nearly double the population of Arizona is largely Republican at present The Republicans in their rule will provide that no amendment shall be con sidered

I THE T MEs71 I world Fair Contests it OFFEH NO lTf acid the three employes of the District or National + t tional Government collecting respectively the < Uteat number ofLouis Sti 4 Louis Exposition coupons to the Worlds Fair for 4 one week and payaIixpenses i pxpenses Note District or National Government ewtploUli es SUKonly Uli only the coupon

Fig. 2. Samples of relatively good (above) and poor (below) OCR from the 1904 *Washington Times*

2 Segmentation

The goal for the first step in the classification process is to identify and categorize meaningful segments of the newspapers. Across NDNP projects and collections, the METS Alto files differ in the detail with which they identify regions, thus we explored extracting segments from the OCRd text; however, none capture smaller segments such as classified advertisements. Segmentation is challenging because of content which varies widely in size and shape, sometimes even wrapping around pictures. There has been some previous work on segmentation (also called zoning) that identified many spatial factors and other segmentation issues [5, 6]. The OCR text can contribute provide semantics.

2.1 Method

Because we have OCRd text with word coordinates, we decided to use methods which incorporated that information. Three techniques of article segmentation were explored. Technique 1 relied on the identification of news article titles determined by several lines of capital letters. In Technique 2, a latent semantic indexing (LSI)-based linear segmentation technique [4] was used to divide the OCR text into blocks. The edges of the blocks were further identified by the closest lines of capital letters which might indicate titles. With Technique 2, the 5594 pages of the *Washington Times* in 1904 yielded 78297 regions. This approach seems to work well for simple segments but not for complex segments such as those split across several columns. Moreover, due to the poor quality of the OCR text, capital letters in the titles of many news articles were originally incorrectly identified, which caused many errors in determining block boundaries. In Technique 3 a combination of approaches was applied. The ratio of capital letters in each text line and the average font size of the titles (which should be larger than the body text) were used to developed a more accurate title detector. If the news articles are located in either the same block or in different blocks of one column, the text blocks are separated into regions by these titles. If the news articles fall across blocks in several columns, these text blocks are linked with the nearest titles that have smaller vertical coordinates compared to the vertical coordinates of the first text lines of the blocks.

2.2 Results

In Technique 3, the pages of the *Washington Times* for 1904 are divided into 116964 regions, an average of about 320 per day (though the Sunday was much higher). The number of segments identified is larger than with the other techniques apparently because it was better at distinguishing short items such as notices and advertisements.

We focused on the performance for one day, picked at random, April 5, 1904. Figure 3 shos images from pages 1 and 3 for that day with segments indicated. As can be seen from inspection of the images, the segmentations are generally accurate. Table 1 divides the results into different types of errors for the first four pages. 67% of the segments were either totally correct or involved only minor errors. Some of the errors affected major articles. For example, the segmentation of the very first article (at the top left of Figure 3) included the newspaper banner but then also incorrectly segmented the lower portion of that article. In the middle of that page, a segment was

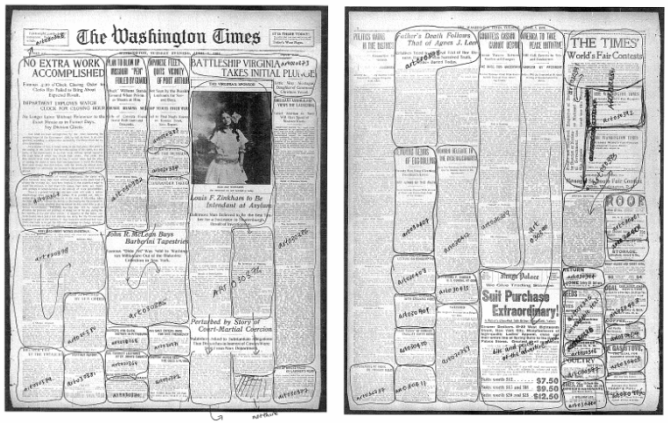

Fig. 3. Pages 1 and 3 of the *Washington Times* for April 5, 1904 with extracted regions

missed. The caption of the picture on that page was also incorrectly joined to the article below it. For smaller items such as classifieds and notices, the most common error was merging text blocks, but these are often difficult for a human to parse without careful inspection. While the program performed well there were some highly complex headings where it continued to have difficulties. Similarly, it seemed to have particular difficulty with the captions of images and cartoons. Furthermore, spot checking its performance suggests that the performance we report for April 5 is typical.

Table 1. Summary of the results for the newspaper segmentation for the first four pages for April 5, 1904

	Count	*Percent*
Correct	64	67
nor errors (e.g., merging a few words)	6	6
Combined two or more articles	11	12
Too much segmentation	14	15

3 Genre Identification

As suggested by the second step in the classification process of the pipeline model, we need to identity the various types of content. This analysis aims to separate the news stories from the wide variety of other material in the newspaper. Some of those contents will be of interest in themselves. Indeed, several NDNP participants have paid to have the obituaries extracted manually because they are of considerable interest for genealogists. Potentially a technique such as ours could automate that process. In other cases, the articles can be analyzed with topic categories and event categorization. The International Press and Telecommunications Council (IPTC) is a newspaper trade organization which has proposed controlled vocabularies for news genres and subjects. We favored these because the terms were designed for news publishing and are used for many modern newspapers (cf., [3]). Even within a specific newspaper, there are often substantial changes in layout and types of coverage across time. While it would

be possible to tune the parameters of the program to this particular newspaper to improve the accuracy of the results, our goal is create a robust program.

3.1 Method

We used the IPTC genres as the basis for our categories, but applied the idea of genres somewhat loosely. Our primary goal is to identify different types of content and extract them. Moreover, the IPTC genres include some which are likely to be quite difficult to identify by automatic methods such as separating "background" from the story itself, and omitted other genres such as advertisements. Thus, we added other categories matching what we found in the *Washington Times*, including a set of advertisement sub-genres and game categories such as chess. Table 2 shows an example of genres and sub-genres.

Table 2. The genre categories used in this analysis

ads:autos	ads:travel	congressional notes	lost_found	sports:golf
ads:clothes	ads:whiskey	financial:securities	masthead	sports:horse racing
ads:hats	advertised letters	foreign mails	notices:auctions	sports:tennis
ads:insurance	banner	fraternal organizations	notices:church	sports:track
ads:medicines	bids solicited	games:bridge	notices:death notice	transporation-shipping
ads:paint	classifieds:helpwanted	games:chess	notices:music	vital records
ads:palmistry	commodities:cotton	legal notices: general	notices:navy	weather reports
ads:shoes	commodities:grain	legal notices:probate	obituary	

The primary technique for identification was based on matching words associated with the genre. A sample of terms used for some of the genre categories is shown in Table 3. In some cases, we also used phrases such as "for sale". In addition, we used newspaper and region-specific terms. For example, for DC we included geographic terms such as "Alexandria", "Bethesda", and "Potomac", and the names of the local baseball team, "The Senators".

Table 3. Terms for some specific genre categories

Ads: Medicines	drugs, cure, cures, liver, kidneys, prescriptions, drug, pains, blood, nervous, eye, pain, dying, bone, extract, potency, ache, brain, skin, rectum, chronic, tonic, stomach, remedies, constipation, bottle, bladder, medicine, pills
Chess	chess, check, checkmate, mate, pawn, rook, game, match, win, problems, capture
Weather report	weather, report, temperature, degrees, rain, sun, snow, warmer, colder, temperatures, cloudy, icy, rainy

A score was calculated for each genre for each segment.[1] This score was compared to a genre-specific minimum threshold; typically, that threshold was in the range of 0.002 to 0.010. Other clues were also found to be useful; for example, some of the items had a lot of numbers (e.g., stock market tables, train schedules) in them and others had lists of names (e.g., advertised letters). Thus, we developed counters for

[1] $\dfrac{\left(\dfrac{\#\text{occurrences of all terms}}{\#\text{different terms}}\right)}{\text{length of the segment}}$

those and used those counters to augment the scores. The segment was assigned to the genre which had the highest score.

3.2 Evaluation

For this evaluation we focused on the segments from April 1904. Overall, the program identified and tagged the genre of about half of the total number of segments. These could be separated from the untagged segments. Many of those untagged segments were news stories which should be passed to the following stage of the pipeline for subject categorization. Clearly, not all of these segments were articles. Some of them were very short segments which were essentially incomprehensible. In addition, we continued to find categories which did not match well the categories in Table 3, such as advertisements for beers or advertisements for hens. It is difficult to decide where the subdivisions should stop. For example, we found a gradation of segments which fell between prose articles about baseball and segments which were composed primarily of statistics. Detailed results for some of the narrower genres are shown in Table 4. Performance is measured in terms of precision and recall and is based on the ratings of a human judge who is considered to be an expert. Precision is the number of articles correctly judged by the program to be in the genre divided by the total number judged in that category (including the ones incorrectly identified). Recall is the number of articles correctly identified by the program as matching the genre divided by the total number of articles judged by the expert to be in that genre. Precision and recall scores are common range from 0.0 to 1.0. Out of a total of 1028 articles that contained words that are in the weather genre, 30 were actual weather report files. The program identified 21 of those correctly, missed nine, and misidentified two. Note that weather reports were sometimes confused with genres such as the crop report which also mentioned the effects of weather on the crops but did not contain a specific weather report. Specific techniques could have been adopted to improve performance still further. For example, the word “weather” generally appeared near the beginning of the weather reports and we could have built a detector specifically for that. However, we were not sure that would be generally applicable for other newspapers so we did not implement it here.

Table 4. Accuracy for selected categories. Narrow genre categories such are recognized quite well.

	Precision	*Recall*
Advertised Letters	0.95	0.74
Games: Chess	0.75	0.59
Weather reports	0.91	0.70

4 Conclusions

4.1 Extensions of the Segmentation Analysis

There are three areas where we believe the segmentation procedure could be productively extended. First, the segmentation algorithm seems to have been confused by pictures and drawings because they leave a hole where there is no text. Automated analysis of the layout of the page-image file should be conducted to identify the locations of pictures and drawings so they can be accounted for in the segmentation process. Second,

some of the segments which were difficult to identify in the genre program were a combination of several unrelated notices. Presumably, this was because they did not include distinct headings but, perhaps, they could be disentangled with semantic analysis. We also need to explore the generality of the results across different years and different newspapers. Finally, as noted earlier Alto files from different digitization projects include different levels of segmentation information. Sometimes that includes keystroked article headings. Additional tests should be conducted to determine how useful that information would be in aiding accurate segmentation.

4.2 Extensions of Genre Identification and Adding Feedback to the Pipeline

Compared to the segmentations, quite a lot of improvement should be possible for genre identification. While we have already incorporated a number of factors, there are still other constraints which should be explored. Here, we examine two possibilities. First, in some cases such as weather reports we know there should be one and only one instance each day. Such a constraint could be strictly enforced. Second, we plan to add a named-entity identification utility. Names such as those of government officials would help pin down the nature of the segment which mentions them. In addition, as with the segmentations, we need to explore the generality of the genre identification results for different years and for different newspapers. More ambitiously, we have obtained the best results when using specific genres (e.g., ads:medicine) rather than general ones (e.g., literature or opinion). Although it may be fairly subtle, the text for those general categories should be distinctive and might be able to be processed automatically. We plan to apply and test several machine learning techniques (e.g., [9]) for categorization. The Pipeline Model has proven to be an effective basis for initial analyses, but for further analyses, it seems that feedback should be helpful. For example, the genre and subject categories could improve the original term disambiguation. Similarly, using known headings to look for segment headings such as "Special Notices" could be used to improve the segmentation.

4.3 The Big Picture

We will be continuing to the next stages of the pipeline; that is to subject categorization and event identification. Indeed, we are exploring advanced strategies for categorization [9]. In addition to genre codes, the IPTC also provides a controlled vocabulary for subjects (topics) and we will use them. As with genre analysis, there are many factors which contribute to accurate subject categorization. In addition to named entities, there are temporal patterns which could be exploited in an analysis program (Table 5). The year 1901 was selected because it covers the death of President McKinley and the beginning of Theodore Roosevelt's presidency.

Table 5. Month-by-month word counts for selected terms in the *Washington Times* for 1901. For the term "drought" there is a clear seasonal pattern. For other terms such as "President" and "Roosevelt" there was a sharp increase in the count in September when he became president.

Terms	*J*	*F*	*M*	*A*	*M*	*J*	*J*	*A*	*S*	*O*	*N*	*D*
Drought	4	6	4	3	8	7	53	21	10	3	7	3
President	877	746	984	811	787	762	460	358	1798	856	932	1070
Roosevelt	37	50	70	14	13	8	7	38	198	201	200	222

More generally, we believe that providing explicit knowledge about history and about the newspaper will be helpful. However, it does not seem feasible to do that by entering specific facts; there are just too many. Generative models such as cyclic models for the seasons may be better (see [2]). These could be simple such as listing the months of the baseball season, the years in which there are presidential elections, the years during which the Wright Brothers were working, or the locations of major buildings in the city. Beyond categorization, we would like to consider other text-based services such as summarization and the identification of specific events. Doing that would make it much easier to develop timeline interfaces (cf., [1]) and to link to resources such as Wikipedia. Clean text would be very helpful for those services and hopefully, we will be able to use the categories to help correct the text. However, it seems unlikely that automated text correction can yield very high accuracy. To reach that level of accuracy is likely to require some sort of human intervention. Perhaps if the quality of the text is relatively high, the corrections could be made by professionals, but it might also be possible to recruit members of a local historical society to make the changes (cf., [8]).

Acknowledgements

This work was supported in part by an NEH Digital Humanities Start-up Research Grant. We thank Abhijeet Ganachari for assistance and we thank Ray Murray of the Library of Congress for the *Washington Times* files.

References

1. Allen, R.B.: A Focus-Context Timeline for Browsing Historical Newspapers. In: ACM/IEEE Joint Conference on Digital Libraries, pp. 260–261 (2005)
2. Allen, R.B., Japzon, A., Achananuparp, P., Lee, K.-J.: A Framework for Text Processing and Supporting Access to Collections of Digitized Historical Newspapers. In: HCI International Conf. (2007)
3. Allen, R.B., Schalow, J.: Metadata and Data Structures for the Historical Newspaper Digital Library Project. In: ACM CIKM, Kansas City, November, pp. 147–153 (1999)
4. Choi, Y.Y.: Advances in domain independent linear text segmentation. In: Proceedings of NAACL, Seattle, USA (2000)
5. Gatos, B., Gouraros, N., Mantzaris, S., Perantonis, S., Tsigris, A., Tzavelis, P., Vassilas, N.: A New Method for Segmenting Newspaper Articles. In: SIGIR, p. 389 (1998)
6. Kanungo, T., Allen, R.B.: Full-Text Access to Historical Newspapers. Technical Report: LAMP-TR-033/CAR-TR-915/CS-TR-4014, University of Maryland, College Park (April 1999)
7. Murray, R.: Towards a Metadata Standard for Digitized Historical Newspapers. JCDL, 330–331 (2005)
8. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: ReCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science 321, 1465–1468 (2008)
9. Zhu, W., Allen, R.B.: Topic and Event Categorization of Historical Newspapers (in preparation)