Faculty Publications

2012

A Synthetic Document Image Dataset for Developing and Evaluating Historical Document Processing Methods

Daniel Walker

William Lund

Brigham Young University - Provo, billlund@icloud.com

Eric Ringger

Follow this and additional works at: <https://scholarsarchive.byu.edu/facpub>

Part of the [Archival Science Commons](#), and the [Computer Sciences Commons](#)

Original Publication Citation

Walker, D. D., Lund, W. B., and Ringger, E. K. (2012) A Synthetic Document Image Dataset for Developing and Evaluating Historical Document Processing Methods. In Proceedings of Document Recognition and Retrieval XIX (DRR 2012), San Francisco, Calif. DOI: 10.1117/12.912203

BYU ScholarsArchive Citation

Walker, Daniel; Lund, William; and Ringger, Eric, "A Synthetic Document Image Dataset for Developing and Evaluating Historical Document Processing Methods" (2012). *Faculty Publications*. 1649.
<https://scholarsarchive.byu.edu/facpub/1649>

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

A Synthetic Document Image Dataset for Developing and Evaluating Historical Document Processing Methods

Daniel Walker, William Lund, and Eric Ringger

Natural Language Processing Lab, Computer Science Dept.
Brigham Young University, Provo, UT, USA

ABSTRACT

Document images accompanied by OCR output text and ground truth transcriptions are useful for developing and evaluating document recognition and processing methods, especially for historical document images. Additionally, research into improving the performance of such methods often requires further annotation of training and test data (e.g., topical document labels). However, transcribing and labeling historical documents is expensive. As a result, existing real-world document image datasets with such accompanying resources are rare and often relatively small. We introduce synthetic document image datasets of varying levels of noise that have been created from standard (English) text corpora using an existing document degradation model applied in a novel way. Included in the datasets is the OCR output from real OCR engines including the commercial ABBYY FineReader and the open-source Tesseract engines. These synthetic datasets are designed to exhibit some of the characteristics of an example real-world document image dataset, the Eisenhower Communiqués. The new datasets also benefit from additional metadata that exist due to the nature of their collection and prior labeling efforts. We demonstrate the usefulness of the synthetic datasets by training an existing multi-engine OCR correction method on the synthetic data and then applying the model to reduce word error rates on the historical document dataset. The synthetic datasets will be made available for use by other researchers.

Keywords: synthetic document images, OCR, datasets, document degradation models, historical document processing

1. INTRODUCTION

A document image dataset that includes gold-standard transcriptions and optical character recognition (OCR) output can be useful in several types of document processing research. The most obvious work that can benefit from such a dataset is the investigation of improved OCR systems or OCR error correction algorithms. In this case, a gold-standard transcription is required in order to assess the effectiveness of the methods being evaluated by the researcher. If the labeled dataset is similar to real-world datasets then these types of evaluations on transcribed datasets can help practitioners choose the best algorithms to use when OCRing new document images for non-transcribed data.

Other types of research may not be concerned with correcting errors but with the impact that these errors have on performance in text processing and retrieval tasks.¹⁻³ Once performance has been assessed, the data may then be used to help build document indexing, summarization, topic modeling or parsing algorithms that are robust to the presence of these errors. In these cases, the researcher would like to be able to evaluate the methods in question against various datasets with varying amounts of noise, in order to assess the degree to which OCR errors degrade performance on the target task.

In all cases, it is desirable to have common datasets that researchers can use as benchmarks in their research. A dataset that includes images, OCR output and gold-standard images together allows researchers to have a common starting point for their methodology. If any of these components is not present, then it becomes difficult to compare results produced by distinct groups. For example, if reference OCR text for the images is not provided, it is difficult to know whether differences in final results arise due to the methods under investigation, or to differences in the OCR engine versions or configurations.

There are a small number of historical document image datasets for which reference OCR text and gold-standard transcriptions exist. These are produced at relatively high cost, and are typically quite small in scale compared to datasets that a practitioner would encounter in real-world scenarios. One example of a historical document dataset is the Eisenhower Communiqués,^{4,5} a collection of 610 facsimiles of typewritten documents issued by the Supreme Headquarters Allied Expeditionary Force (SHAEP) during the last years of World War II. Having been typewritten and duplicated using carbon paper, the quality of the print is poor, making them a challenging case for OCR engines. The communiqués have been OCRed using five OCR engines: ABBYY FineReader for Windows version 10,⁶ OmniPage Pro X for Mac OS X,⁷ Adobe

Acrobat Pro for Max OS X,⁸ ReadIris Pro for Mac OS X,⁹ and Tesseract version 1.03.¹⁰ A manual transcription of these documents serves as the gold standard from which it is possible to obtain accurate measures of average document word error rates (WER) for each engine, where a single document WER is defined to be:

$$WER = 100 * \frac{\text{insertions} + \text{deletions} + \text{substitutions}}{\text{total tokens in document}}$$

which can exceed 100% when the numbers of insertions are high. See Table 1 for the average document WERs for the Eisenhower Communiqués data.

OCR Word Error Rates					Average WER
ABBYY	Omnipage	Adobe	ReadIris	Tesseract	
18.2%	30.0%	51.8%	54.6%	67.8%	44.5%

Table 1: Word Error Rates of the five OCR engines used on the Eisenhower Communiqués

Although this dataset has many desirable features, it is small. This means that there is not enough data to build useful training and test sets for many tasks. The text is also very homogeneous in vocabulary and subject matter and is thus not well suited for research involving topic modeling or document clustering on noisy historical documents. Nonetheless, the image quality and OCR error rates are representative of the problem one encounters when working with facsimiles of historical documents and illustrate the need for robust text processing and retrieval techniques when working with historical data.

We set out to create a set of synthetic datasets that would be useful in as many of the above mentioned research scenarios as possible. Our goal was to produce a set of datasets that satisfies the following requirements, each inspired by research needs:

1. The datasets should contain the same data at various levels of degradation.
2. The datasets should be reasonably large, containing thousands of documents each.
3. Each document should have a human-provided topical label, to enable research in noisy text analytics.
4. The datasets should be heterogeneous in the amounts and kinds of noise they contain, so as to be consistent with trends observed in real-world historical documents.
5. Most importantly, the errors in the synthetic data should be as like the errors produced by an actual OCR engine on actual degraded documents as possible.

Requirements 1-3 were met by selecting as the source data three datasets commonly used in the document classification and clustering research literature: 20 Newsgroups,¹¹ Reuters 21578,¹² and the LDC-annotated portion of the Enron e-mail archive¹³ *. Requirements 3 and 4 were met by rendering the digital text documents to images, and then corrupting the images using a parameterizable document degradation model (see Section 3) from the literature with stochastically chosen parameters (see Section 4), and then OCRing the resulting images using the ABBYY and Tesseract OCR engines.

Our contributions include: 1) the synthetic datasets themselves, which we will be releasing and distributing, 2) the methodology and code used in the creation of the datasets, which will also be released with the data, and 3) the evaluation and verification of the datasets which we conducted in order to verify that it is useful for OCR error correction research.

The remainder of the paper is organized as follows. In Section 2 we describe related work to this research. In Section 3 we give an overview of Baird’s document degradation model, including its parameters and the effect they have on the degraded image. Section 4 describes how the document degradation model was used to create datasets with increasing average word error rates consisting of documents with heterogeneous word error rates. Section 5 gives relevant statistics for the synthetic datasets, with comparisons to the statistics of the Eisenhower Communiqués. Section 6 provides a practical example of how the data can be used for model training in an existing OCR error correction model.¹⁴ Finally, Section 7 summarizes the paper and presents our conclusions.

*Due to our agreement with the LDC, only the raw corrupted data, and not the topic annotations from the LDC are distributed with our Enron datasets.

2. RELATED WORK

Document degradation models have been studied extensively in the literature.¹⁵ Perhaps the best known and well regarded being the work done by Henry Baird and his colleagues.¹⁶ We use a two-parameter version of the model described in Sakar, Baird, and Zhang,¹⁷ which will be explained in more detail in Section 3.

Other synthetic document image datasets exist as well. In 2008, Daniel Lopresti introduced a dataset that he produced from the Reuters21578 dataset by physically printing the digital documents with a Ricoh Aficio digital photocopier and then scanning these printouts with the same machine.² Each document was scanned five times. The first was a scan of the original print of the document, another scan was produced by setting the machine to its darkest contrast setting, and another by scanning that output. This procedure was then repeated on the lightest contrast setting to produce the other two scans. The scanned images were then OCRed using the Tesseract OCR engine.

The Lopresti dataset is a valuable resource for researchers in need of noisy text data and can be used to demonstrate the effect of noise on the performance of NLP algorithms for many tasks.² It has the advantage that, although it is synthetic, it makes use of actual physical, optical and mechanical systems in order to degrade documents. However, the dataset does not meet some needs. For example, the dataset consists of 3,305 total pages. However, because each original document is duplicated in the dataset 5 times, there are only 661 unique documents. Consequently, the dataset is too small for realistic research for text analytics and it therefore does not satisfy Requirement 2 from Section 1. Also, it turns out that even for the most degraded documents in the dataset—the second-generation copies made with extreme contrast settings—the documents were not sufficiently degraded to produce a wide range in average word error rates to match those found on real-world historical documents. This shortcoming means that the Lopresti dataset does not satisfy our Requirements 1 and 4 either.

3. DEGRADATION MODEL

The model we chose for document degradation is a simplified version¹⁷ of Henry Baird’s full optical degradation model.¹⁶ The model is controlled by two parameters, the blur standard deviation b , and the binarization threshold t . The degradation of a document is conducted as follows, given a document that is originally created or sampled at a higher resolution:

1. Translate the image randomly uniformly in x and y between 0 and 1 equivalent output pixels. That is, if the image has been rendered at 5 times the output resolution, then translate between 0 and 5 original pixels. This introduces variation due to pixel alignment variations.
2. Use a Gaussian convolution kernel with standard deviation b to blur the image.
3. Subsample the image to the output resolution.
4. Model image sensor pixel sensitivity error by adding a random value to the intensity at each pixel according to a Gaussian distribution centered at 0.0 with a standard deviation of 0.025.
5. Threshold the image to produce a bi-level output image such that all pixels with intensity less than t become 0.0 (black) and all pixels with an intensity greater than or equal to t become 1.0 (white).

4. METHODOLOGY

The first step in producing a degraded dataset is rendering the digital text from the source data to images to be used as input for the OCR process. We chose to use the $\text{\LaTeX}$ document typesetting system for rendering text to images. This choice was motivated by the fact that, with a relatively small amount of cleaning and marking up the source data, $\text{\LaTeX}$ is able to render text using a complex set of rules that results in high quality text layouts, including ligatures, kerning, and automatic pagination when needed.

The output from $\text{\LaTeX}$ is a document in the Adobe Portable Document Format (PDF). In order to obtain an image from this file, we use Ghostscript, a program available on most Linux distributions, to render the PDF to a bitonal Tagged Image Format (TIFF) image at 1500 dpi. The TIFF images are then passed through the document degradation model to produce the noisy document images.

Creating datasets with one setting for the document degradation model, whether that be a single contrast setting on a copier or a single choice of b and t parameters using the model described in Section 3, produces documents with very uniform characteristics. Doing so is at odds with observations taken on the Eisenhower dataset, which reveals that the

documents in real-world historical data collections can vary significantly in terms of quality from one to another. For example, some documents may have been reproduced under less than ideal circumstances. Others may have experienced physical damage. Variations can even occur during the original production of the documents. For example, some of the Eisenhower Communiqués seem to have been typed on machines with fading ink ribbons or were typed by inferior typists whose only options for correcting typing mistakes was to type over mistaken characters with the correct ones.

These differences are reflected in the variation in WERs that are observed on the Eisenhower data, as shown in Figure 1. It can be seen in 1c and 1d that even the Tesseract engine, which had an average WER of 50.1% on this data contained many documents that had relatively low WERs. In fact, for the Tesseract engine, only 29% of the documents had a WER greater than 50% and 43% of the documents had WERs less than 20%.

It would be difficult or impossible to model all of the ways that an historical document can be degraded. However, we decided that, to be true to our real-world data the synthetic OCR documents should consist of documents at various levels of degradation. Thus the distinction between a dataset with low average WER and one with high WER is not that all documents in the high WER dataset have higher WERs. Both datasets consist of a mixture of documents with low, medium, and high amounts of degradation, but the high average WER dataset consists mostly of medium and high WER documents, and the low WER dataset consists mostly of low and medium WER documents. In order to achieve this heterogeneity, we

Figure 1: (a)-(c) show histograms of the WERs observed on the Eisenhower data using three OCR engines. (d) shows the Tesseract WERs, omitting those greater than 100, in order to show more detail in the 0-100% range.

decided that documents should be degraded with randomized choices for the b and t parameters. We chose to parameterize this process with a single parameter α such that higher values of α would produce datasets with documents that are more degraded on average than documents in datasets with lower α values.

Figure 2: The relationship of the t and b parameters on average word error rate of OCR'd documents.

Figure 3: Samples of (b, t) parameter pairs at four different α values. Fifty points were drawn for each α value.

Sakar et al.¹⁷ show the relationship between the t and b parameters and the resulting image quality. In that work, they show how, for a fixed value of t increasing b results in degraded text for which OCR results contain more errors. In addition, for a fixed value of b , there is a “sweet spot” t_s for the t parameter at which documents are degraded in such a way that minimizes OCR error rate, and error rates increase as $|t - t_s|$ increases. As t gets smaller, the lines that form the letter glyphs grow fatter, filling in the gaps in letters such as “a”, “o”, and “R”, and joining letters. As t gets larger, the lines grow thinner, causing single letters to spit, as their thinnest parts are eliminated.

Figure 2 shows how word error rates vary over values of the b and t parameters. This graph was produced by first randomly choosing 10 random documents and then sweeping the 10x10 grid of values for b between 1.3 and 2.65 in increments of .13 and values of t between .1 and .4 in increments of .03. The documents were degraded with each pair of b and t in this grid and the degraded documents were OCR'd using ABBYY FineReader. For the documents degraded with the same parameter values the resulting word error rates were averaged, and that average word error rate is plotted in the z direction. It can be seen that, in general, as b increases, average WER also increases. In addition, for any given b value, average WER increases as $|.20 - t|$ increases.

As mentioned earlier, the degradation model parameter generation process was controlled with a single parameter α . Using α , we draw the t and b parameters for a document according to the following distributions:

$$b \sim (2.6 - 1.3)\text{Beta}(\alpha, 1.0) + 1.3$$

$$t \sim \text{Norm}(.225, ((\alpha/11.0) * (.25 - .001)) + .001)$$

The first distribution is a Beta distribution, scaled and translated to the interval (1.3, 2.6). The second distribution is a Normal distribution with mean .225 and standard deviation that increases linearly from .001 to .25 as α ranges from 0 to 11.0. Figure 3 plots (b, t) points drawn from these distributions at four values of α in order to illustrate how the distributions of b and t vary jointly as α changes. As α increases, sampled b values tend to be greater. Also, though the distribution in t values remains centered at .225, the variance increases, increasing the number of points with t values far from the mean.

Using this process, for each source dataset and each α value in 0.1, 1.0, 2.0, . . . , 10.0 we drew a random parametrization (b, t) for each document based on the current α and then degraded that document. This resulted in 33 datasets, one for each $(\alpha, \text{source dataset})$ pair.

5. STATISTICS

The procedure discussed above produced 452,726 degraded image pages: 239,326 for 20 Newsgroups, 123,748 for Reuters and 89,652 for Enron. Figure 4 shows how the ABBYY FineReader OCR engine performed on these images. This outcome demonstrates that we have created datasets with varying levels of noise, as measured by the average word error rates of the

Figure 4: Correlation between α corruption levels and the resulting average word error rates from ABBYY FineReader.

constituent documents. This outcome also verifies that our synthetic data satisfy Requirement 1 from Section 1. Figure 5 shows histograms of word error rates for the 20 Newsgroups dataset, degraded using four different values of α . Although the histograms appear to be much more regular than the Eisenhower histograms, they do exhibit the desired trend that many documents have low error rates, a noticeable difference being that there is a sharp drop-off at 100%, which does not occur in the Eisenhower data. Despite these small differences, these graphs give some evidence that we have also met Requirements 4 and 5. Further evidence to this effect will be given in the form of an OCR correction task in Section 6.

6. OCR ERROR CORRECTION TASK

In order to further validate the utility of the synthetic datasets, we used the data to train an existing approach to multiple-engine OCR error correction, as described by Lund et al.¹⁴ The method is motivated by the reasoning that OCR engines have different strengths and weaknesses. Thus, if one OCR engine outputs an incorrect hypothesis for a word token in the source image, another engine might output the correct hypothesis. The problem then becomes, given the output from multiple OCR engines based on the same source image, to choose between the hypotheses at each word in such a way so as to minimize the word error rate of the final output.

We employ a Maximum Entropy multi-class classifier¹⁸ as implemented in the MALLET toolkit,¹⁹ trained on a subset of the synthetic data introduced here, consisting of 785 documents from the Enron data at three different α levels: 1.0, 5.0, and 9.0. We call these three datasets the *calibration sets*. We prepared training data from the synthetic document image datasets by first aligning the output of the OCR engines in order to produce aligned columns where each column roughly corresponds to hypotheses for the same word in the source image. We then extracted the following features from each column:

Voting: indicates when multiple hypotheses in a column match exactly,

Number: binary indicators for whether each hypothesis is a cardinal number,

Dictionary: binary indicators for whether each hypothesis appears in the Linux dictionary,

Gazetteer: binary indicators for whether each hypothesis appears in a gazetteer of place names, and

Spell Checker: an additional hypothesis generated by Aspell (citation) from words that do not appear in the dictionary or in the gazetteer.

For each training case (an aligned column), the label indicates which OCR engine provided the correct hypothesis. Ties were resolved by selecting the output from the OCR engine with the lowest overall WER on the training data (ABBYY FineReader).

We trained individually on the three calibration sets, and used the resulting classifiers to choose the transcription hypotheses for the Eisenhower data, given the output for that dataset from the two OCR engines, for which OCR output has been included in the datasets. The results are shown in Table 2. Recall from Table 1 that the best single OCR engine

(a) $\alpha = 0.1$ (6.8% average WER)

(b) $\alpha = 1.0$ (10.3% average WER)

(c) $\alpha = 5.0$ (31.7% average WER)

(d) $\alpha = 10.0$ (41.2% average WER)

Figure 5: Histograms of the word error rates of ABBYY FineReader on the 20 Newsgroups synthetic datasets at four α levels. The x-axis has been truncated to the (0,200) range.

Callibration Set (α) Used for Training	Resulting Average WER on the Eisenhower dataset
1.0	16.73%
5.0	15.56%
9.0	22.87%

Table 2: Average word error rates of the final output of multi-engine error correction algorithm trained on the Enron calibration sets Eisenhower Communiqués

(ABBYY) achieved an 18.2% WER. The improvements shown in rows 1 and 2 indicate that these datasets are useful for training the correction model in a way that improves accuracy on the Eisenhower data. This result is significant because it tells us that this synthetic data can be used to train models for improving OCR output on un-transcribed historical document images. If the synthetic data consisted of errors that were unlike those seen in real-world data, then the patterns learned during training on the calibration sets would not have generalized to the Eisenhower data, precluding this level of improvement. We believe this is additional evidence that the synthetic datasets are sufficiently like real-world historical OCR data to be useful for OCR research. (Our previously published experiments showed the value of this multi-engine error correction method, reducing word error rates to 13.29% – a 26% relative reduction in WER – by adding the output from three additional OCR engines.¹⁴)

7. CONCLUSIONS

We have introduced a set of synthetic datasets of degraded document images together with gold standard text and the output from two OCR engines. The data have nice statistical properties in terms of the word error rates observed with the OCR engines. In addition, we have presented evidence that the data are sufficiently like real-world OCR data.

Though the document degradation model we used has been introduced before, our method for stochastically choosing parameters per image in order to simulate differences observed in real historical data is novel, and the methodology could be applied to other source data.

For instructions on how to download the synthetic datasets, the code used to produce it, the Eisenhower Communiqués, and document image samples please visit:

https://facwiki.cs.byu.edu/nlp/index.php/Synthetic_OCR_Data.

REFERENCES

- [1] Lopresti, D., “Performance evaluation for text processing of noisy inputs,” in [*Proceedings of the 20th Annual ACM Symposium on Applied Computing (Document Engineering Track)*], 759–763 (Mar. 2005).
- [2] Lopresti, D., “Optical character recognition errors and their effects on natural language processing,” in [*AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*], 9–16, ACM, New York, NY, USA (2008).
- [3] Daniel D. Walker, William B. Lund, E. K. R., “Evaluating models of latent document semantics in the presence of OCR errors,” in [*Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*], (2010).
- [4] Jordan, D. R., “Daily battle communiques, 1944-1945,” Harold B. Lee Library, L. Tom Perry Special Collections, MSS 2766, <http://lib.byu.edu/digital/spc/eisenhower> (1945).
- [5] Lund, W. B. and Ringger, E. K., “Improving optical character recognition through efficient multiple system alignment,” in [*Proceedings of the Joint Conf. on Digital Libraries (JCDL'09)*], (June 2009).
- [6] ABBYY, “ABBYY finereader,” <http://finereader.abbyy.com> (2010).
- [7] Nuance Communications, Inc., “OmniPage Pro,” <http://www.nuance.com/imaging/products/omnipage.asp> (2010).
- [8] Adobe Systems Inc., “Acrobat pro,” <http://www.adobe.com/products/acrobatpro.html> (2010).
- [9] I.R.I.S. s.a., “Readiris pro,” <http://www.irislink.com> (2010).
- [10] Google, Inc., “Tesseract,” <http://code.google.com/p/tesseract-ocr> (2010).
- [11] Lang, K., “NewsWeeder: Learning to filter netnews,” in [*Proceedings of the Twelfth International Conference on Machine Learning*], 331–339 (1995).
- [12] Lewis, D., “Reuters-21578 text categorization test collection,” <http://www.research.att.com/~lewis> (1997).
- [13] Berry, M. W., Brown, M., and Signer, B., “2001 topic annotated Enron email data set,” (2007).
- [14] Lund, W. B., Walker, D. D., and Ringger, E. K., “Progressive alignment and discriminative error correction for multiple OCR engines,” in [*Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011)*], (Sept. 2011).
- [15] Baird, H., “The state of the art of document image degradation modelling,” in [*Digital Document Processing*], Chaudhuri, B. B., ed., *Advances in Pattern Recognition*, 261–279, Springer London (2007). 10.1007/978-1-84628-726-8_12.
- [16] Baird, H., “Document image defect models,” in [*Document image analysis*], 315–325, IEEE Computer Society Press (1995).
- [17] Sarkar, P., Baird, H. S., and Zhang, X., “Training on severely degraded text-line images,” in [*Proceedings of the IAPR 7th International Conference on Document Analysis and Recognition (ICDAR03)*], (Aug. 2003).
- [18] Nigam, K., Lafferty, J., and McCallum, A., “Using maximum entropy for text classification,” in [*IJCAI-99 workshop on machine learning for information filtering*], 1, 61–67 (1999).
- [19] McCallum, A. K., “MALLET: A machine learning for language toolkit,” <http://mallet.cs.umass.edu> (2002).