2014

How Well Does Multiple OCR Error Correction Generalize?

William B. Lund  
*Brigham Young University - Provo*, billlund@icloud.com

Eric K. Ringger  
*Brigham Young University - Provo*, ringger@cs.byu.edu

*See next page for additional authors*

Follow this and additional works at: [https://scholarsarchive.byu.edu/facpub](https://scholarsarchive.byu.edu/facpub)

Part of the Computer Sciences Commons, and the Library and Information Science Commons

**Original Publication Citation**  

**BYU ScholarsArchive Citation**  
[https://scholarsarchive.byu.edu/facpub/1647](https://scholarsarchive.byu.edu/facpub/1647)

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
How Well Does Multiple OCR Error Correction Generalize?

William B. Lund\textsuperscript{a}, Eric K. Ringger\textsuperscript{b}, and Daniel D. Walker\textsuperscript{c}

\textsuperscript{a}Harold B. Lee Library, Brigham Young University, Provo, UT 84602, USA
\textsuperscript{b}Computer Science Department, Brigham Young University, Provo, UT 84602, USA
\textsuperscript{c}Microsoft, Redmond, WA 98052, USA

ABSTRACT

As the digitization of historical documents, such as newspapers, becomes more common, the need of the archive patron for accurate digital text from those documents increases. Building on our earlier work, the contributions of this paper are: 1. in demonstrating the applicability of novel methods for correcting optical character recognition (OCR) on disparate data sets, including a new synthetic training set, 2. enhancing the correction algorithm with novel features, and 3. assessing the data requirements of the correction learning method. First, we correct errors using conditional random fields (CRF) trained on synthetic training data sets in order to demonstrate the applicability of the methodology to unrelated test sets. Second, we show the strength of lexical features from the training sets on two unrelated test sets, yielding a relative reduction in word error rate on the test sets of 6.52%. New features capture the recurrence of hypothesis tokens and yield an additional relative reduction in WER of 2.30%. Further, we show that only 2.0% of the full training corpus of over 500,000 feature cases is needed to achieve correction results comparable to those using the entire training corpus, effectively reducing both the complexity of the training process and the learned correction model.

Keywords: Historical Documents, Optical Character Recognition, OCR Error Correction, Ensemble Methods

1. INTRODUCTION

Historical machine printed document images often exhibit significant noise, making the optical character recognition (OCR) of the text difficult. Our previous work\textsuperscript{1,2} shows that it is possible for combined outputs from multiple OCR engines using machine learning techniques to provide text output with a lower word error rate (WER) than the OCR of any one OCR engine alone. Further, we use methods which are scalable to very large collections, up to millions of images, without document- or test corpus-specific manipulation of training data, which would be infeasible given time and resource constraints.

Ensemble methods are used effectively in a variety of problems such as machine translation, speech recognition, handwriting recognition, and OCR error correction, to name a few. In a paper on pattern recognition frameworks for ensemble methods, Kittler et al.\textsuperscript{3} state: "It had been observed ... that although one of the [classifiers] would yield the best performance, the sets of patterns mis-classified by the different classifiers would not necessarily overlap. This suggested that different classifier designs potentially offered complementary information about the patterns to be classified which could be harnessed to improve the performance of the selected classifier." Previously we have merged complementary information such as the output of multiple OCR engines\textsuperscript{2} and multiple binarizations of the same document image\textsuperscript{1} on a single test set and training set. The goal of this paper is to demonstrate the generalizability of the methods involving multiple OCR engines, to introduce a new test and a new training set, to show the results of feature engineering, and to demonstrate the degree to which a large training set may be reduced and still yield results consistent with the full training set.

The remainder of this paper proceeds as follows: Section 2 discusses existing work in several fields related to the methods and outcomes of this research. Section 3 outlines a brief overview of the methods used to extract corrected text with machine learning techniques, leading to the heart of the paper: the evaluation of the extent to which these methods are applicable across multiple test corpora and synthetic training sets in Section 4. Finally, Section 5 summarizes the conclusions of this research.

Further author information: (Send correspondence to William Lund)
William Lund: E-mail: bill\_lund (at) byu.edu, Telephone: +1 801 422-4202
Eric Ringger: E-mail: ringger (at) cs.byu.edu
2. RELATED WORK

Extracting usable text from older, degraded documents is often unreliable, frequently to the point of being unusable.\(^4\) Kae and Learned-Miller\(^5\) remind us that OCR is not a solved problem and that “the goal of transcribing documents completely and accurately... is still far off.” At some point the word error rate of the OCR output inhibits the ability of the user to accomplish useful tasks.

Ensemble methods are used with success in this task as well as in a variety of settings. In 1998 Kitter et al.\(^3\) provided a common theoretical framework for combining classifiers which is the basis for much of the work in ensemble methods. In off-line handwriting recognition Bertolami and Bunke\(^6\) use ensemble methods in the language model. Si et al.\(^7\) use an ensemble of named entity recognizers to improve overall recognition in the bio-medical field. For machine translation, Machery and Och\(^8\) present a study in how different machine translation systems affect the quality of the machine translation of the ensemble. Maximum entropy models have been used previously to select among multiple parses returned by a generative model.\(^9\)

Klein and Kobel\(^10\) as well as Cecotti and Belaïdi\(^11\) note that the differences between OCR outputs can be used to advantage. This observation is behind the success of ensemble methods, that multiple systems which are complementary can be leveraged for an improved combined output. The question of how many inputs should be used in an ensemble system is generally “the more the better.” Caruana et al.\(^12\) use on the order of 2000 models built by varying the parameters of the training system to create different models. On a smaller number of inputs (five OCR engines), Lund et al.\(^13\) demonstrate that the error rate of the ensemble decreases with each added system. It should be noted that the complementarity\(^14\) of correct responses of the methods is critical. An important point is that even high error rate systems added to an ensemble can contribute to reducing the ensemble error rate where the addition represents new cases or information not included previously. Diversity in the ensemble is critical to improving the system’s performance over that of any individual in the ensemble.\(^15–18\) This paper will expand on the observation regarding complimentary sources, noting that one useful source of diversity is the output of multiple OCR engines.

Necessary for our post-OCR error correction using multiple sequences is an alignment of the text sequences, which can either use exact or approximate algorithms. The multiple sequence alignment problem has been shown to be NP-Hard by Wang and Jiang.\(^19\) Lund and Ringger\(^2\) demonstrate an efficient means for exact alignment; however, alignment problems on long sequences are still computationally intractable. Much of the work in multiple sequence alignment work is done in the field of bioinformatics, where the size of the alignment problems has forced the discovery and adoption of heuristic solutions such as progressive alignment.\(^20\) Elias\(^21\) discusses how a simple edit distance metric, which may be appropriate to text operations, is not directly applicable to biological alignment problems, which means that much of the alignment work in bioinformatics requires some adaptation for use in the case of text.

It is well known from work by Lopresti and Zhou\(^22\) that voting among multiple sequences generated by the same OCR engine can significantly improve OCR. One practical application of voting can be found in the Medical Article Record System (MARS) of the National Library of Medicine (NLM) which uses a voting OCR server for text recognition.\(^23\) Esakov, Lopresti, and Sandberg\(^24\) evaluated recognition errors of OCR systems, and Kolak, Byrne, and Resnik\(^25\) specifically applied their algorithms to OCR systems for post-OCR error correction in natural language processing tasks. OCR error correction with in-domain training\(^26\) as well as out-of-domain training using a synthetic training dataset\(^13\) have been shown to be effective.

Recent work by Yamazoe et al.\(^27\) effectively uses multiple weighted finite-state transducers (WFST) with both the OCR and a lexicon of the target language(s) to resolve the ambiguity inherent in line- and character-segmentation, and character recognition, in which the number of combinations can be very large. Both conventional OCR and post-processing are contained within their system, resolving the difference between various hypotheses before committing to an output string.

A non-ensemble method for improving historical document images prior to OCR is adaptive binarization. Our previous work\(^1\) compared the OCR results of an ensemble of document image binarizations to the results of adaptive binarization. From the perspective of the corpus WER, the results of the ensemble methods were superior to those of individual document image adaptive binarizations.

A contribution of this paper is an extension of previous methods in supervised, discriminative machine learning methods to choose among all hypotheses, in which previous methods are shown to be effective in two unrelated historical print corpora and two unrelated training datasets. In this work the models are learned on synthetic, out-of-domain training data sets, created and computationally degraded according to the methods proposed by Sarkar, Baird, and Zhang\(^28\) and Baird.\(^29\)
Figure 1. The methodology used in this paper to: create the training set and CRF model, prepare the test set for processing by the CRF, and evaluate the results using the test set transcription.

3. METHODOLOGY

The first step of our methodology prepares the test and training corpora, scanning document images and creating the synthetic training sets. (See Figure 1 for a flow chart of this process.) Once the document images have been scanned, the images are recognized with the selected OCR engines; in this case those are Abbyy FineReader 10, OmniPage 18, Adobe Acrobat Pro X, and Tesseract 3 (an open source OCR system). The baseline OCR results of each OCR engine are seen in
Table 1. Baseline corpus word error rates using micro-averaging for datasets by OCR engine. These results are on the original document images without modification. Note that WERs of greater than 100% are possible due to multiple insertions not found in the reference text.

<table>
<thead>
<tr>
<th>Test Set Corpora</th>
<th>Abbyy Fine-Reader 10</th>
<th>OmniPage 18</th>
<th>Adobe Pro X</th>
<th>Tesseract 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eisenhower Communiqués</td>
<td>19.98%</td>
<td>30.50%</td>
<td>52.59%</td>
<td>93.99%</td>
</tr>
<tr>
<td></td>
<td>Average WER: 49.26%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19th Century Mormon Article</td>
<td>7.44%</td>
<td>11.77%</td>
<td>23.49%</td>
<td>18.35%</td>
</tr>
<tr>
<td>Newspaper Index</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average WER: 15.26%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training Set Corpora</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enron Synthetic Dataset</td>
<td>24.31%</td>
<td>30.57%</td>
<td>68.95%</td>
<td>56.07%</td>
</tr>
<tr>
<td>Reuters-21578 Dataset</td>
<td>15.35%</td>
<td>20.28%</td>
<td>99.77%</td>
<td>82.37%</td>
</tr>
</tbody>
</table>

Figure 2. From “Communiqué No. 1” of the Eisenhower Communiqués. The word error rate of this document across the five OCR engines used in this research varied from 10.63% to 63.41%, with a mean WER of 36.34%.

Table 1. The OCR output is character aligned yielding parallel hypotheses from the OCR engines. Where spaces occur in the aligned texts, the process creates a column of text hypotheses. From these columns, features used by the machine learner are extracted as described in Section 3.3. The machine learner, a trained conditional random field (CRF), labels the aligned hypotheses with the OCR engine to be selected or “NONE”, indicating that no output for the column should be selected. The CRF models are trained using the training sets prepared similarly to the test corpora. Collectively, the text associated with each label from all columns constitute the error corrected output. (See Figure 4 for an example.) The following sections describe in more detail this process.

3.1 Corpora

Four datasets were used in this work: two test sets, the Eisenhower Communiqués and the Nineteenth Century Mormon Article Newspaper Index; and two training sets, an extraction of the 2001 Topic Annotated Enron Email Data Set and an extraction of the Reuters-21578 Text Categorization Test Collection. The following sections describe each dataset and how it was created.

3.1.1 Eisenhower Communiqués

The Eisenhower Communiqués are a collection of 605 facsimiles of typewritten documents created by the Supreme Headquarters Allied Expeditionary Force (SHAEF) during the last years of World War II. Having been typewritten and duplicated using carbon paper, the quality of the print is poor. (See Figure 2 for an example.) A manual transcription of these documents serves as the gold standard for evaluating the word error rates of the OCR. In the course of duplication the documents have effectively become bi-tonal and are treated as such by this research.

---

An online presentation of The Eisenhower Communiqués is viewable at http://www.lib.byu.edu/digital/eisenhower.
3.1.2 Nineteenth Century Mormon Article Newspaper Index

The Nineteenth Century Mormon Article Newspaper Index\(^1\) (19thCMAN) corpus is a collection of 1055 color images\(^1\) of articles dealing with events and persons of The Church of Jesus Christ of Latter-day Saints (Mormon) from an archive of historical newspapers of the 19th century housed at the Harold B. Lee Library of Brigham Young University. As expected from 19th century newprint, the quality of the paper and the print was poor when first printed and has further degraded over time. The newspapers were scanned at 400 dots per inch (dpi) in 24-bit RGB color, and the individual articles were segmented and saved as TIFF images. For previous work the RGB images were converted to 8-bit grayscale. The OCR output of each document was manually corrected by two reviewers to act as a gold standard. An example from the document corpus can be seen in Figure 3.

3.1.3 Synthetic Training Sets

For the training sets, we created synthetic data sets from the 2001 Topic Annotated Enron Email Data Set\(^3\) and the Reuters-21578 Text Categorization Test Collection.\(^3\) From the digital text of each document in the test corpora, a TIFF document image was generated and randomly degraded using techniques inspired by Baird\(^2\) and Sarkar et al.\(^2\) Each synthetic document image was produced using the following steps. First an image is rendered as a bi-tonal document. Spatial sampling error is introduced by translating the entire image stochastically. The image is blurred using a Gaussian convolution kernel. The document is sub-sampled. Gaussian noise is added to simulate pixel sensor sensitivity. To binarize a document, a threshold is applied. For further details on the process, please consult the paper by Walker, Lund, and Ringger (2013).\(^3\)

3.2 Progressive Alignment

Since exact \(n\)-way alignments become exponentially complex in \(n\) we turned to greedy progressive alignment heuristics, which are applied successfully in bioinformatics\(^2\) and textual variance analysis.\(^3\) In brief, progressive alignment algorithms begin by selecting two sequences to be aligned that are most similar based on some similarity measure applied to all sequences. Additional sequences are aligned, using the same selection criteria as for the first two, until all sequences have been aligned. (Refer to Spencer et al.\(^3\) for details on progressive alignment in a textual context.) The order of pairwise alignments is specified in a binary tree structure called the guide tree. Due to downstream consequences of greedy choices, a progressive alignment heuristic is not optimal; however, the resulting alignments are good in practice.

In this paper, the order of the alignment, unless indicated otherwise, is a greedy approximation of the guide tree based on sequence similarity of the training set; specifically in the order: Abbyy FineReader and OmniPage Pro X, then Adobe Acrobat Pro, and lastly Tesseract. The incremental results of this alignment order on the WER can be seen in Table 4.

3.3 Assigning Features in an Alignment Column

We employ modern supervised discriminative machine learning methods trained on the training set. The role of the machine learning model is to select the proper hypothesis from each aligned column in order to produce the best OCR correction. We prepared training data from the training sets with the same OCR engines and aligned their output using the same progressive alignment algorithm described above in order to produce aligned columns of hypotheses. (See Figure 4.) As a base we extracted the following kinds of feature types from each column:

- **Voting**: multiple features to indicate where 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.
- **Lexical Features**: words that appear in the corpus are individually created as features.

\(^1\)An online presentation of The Nineteenth Century Mormon Article Newspaper Index is viewable at http://lib.byu.edu/digital/19cMormonArticles.
Figure 4. An example of an aligned lattice from document CT_2Apr1872_p2_c1. The “dash” character represents a gap or \textit{INDEL} in the alignment, where a character needs to be inserted in order to complete the alignment. Correct hypotheses in the aligned text are underlined. The aligned text is divided into columns of hypotheses on spaces in the aligned sequences. Note that a space occurred in the middle of the Abbyy mis-recognition of the word “Legislature”.

For each training case (an aligned column), the label indicates which OCR engine provided the correct hypothesis. Ties were resolved by selecting the OCR engine with the lowest WER from the training set. Note that “DictA” (and so forth) indicates that the entry from Abbyy is found in the dictionary. Leading and trailing punctuation is removed from the hypothesis before checking in the dictionary. To produce a “Voting” feature type the match must be exact, including punctuation. Once all of the feature vectors have been extracted, we use the maximum entropy learner in the Mallet toolkit to train a maximum entropy (a.k.a., multinomial logistic regression) model to predict choices on unseen alignment columns.

4. RESULTS

Previous work\cite{2,13,26} using the Eisenhower Communiqués test set and the Enron training set was focused on individual document performance in which corpus WERs were calculated as the average of the individual document WERs. A result of this method is that corpus statistics would give more weight to the tokens of a short document than to the tokens of a long document. This approach may be called a “macro-average” in which each document is given an equal weight, regardless of its size. In contrast, the results reported in this paper are based on tokens, giving an equal weight to each occurrence of a token in the corpus, the OCR output, and the resulting error corrections. All averages and other statistics in this paper, unless stated otherwise, use this “micro-averaging” approach. We believe this approach is more suited for feature engineering although there are important uses of a document-by-document evaluation. Examples where a document-by-document analysis is useful would be observing improvement trends by document or in error analysis related to document WER.
Both for documents and for the corpus as a whole, the word error rate is calculated as

\[
\text{WordErrorRate} = \frac{\text{Substitutions} + \text{Insertions} + \text{Deletions}}{\text{Number Of Reference Tokens}}
\]

which may be greater than 100% due to the number of insertions, which is not limited.

The remainder of this section is organized as follows. Section 4.1 shows the baseline results from previous work reinterpreted using the micro-averaging approach discussed above. New features for this work, recurring features and an order 1 CRF, are described in Section 4.2 with the results of incorporating these features in Section 4.3. The generalization of these methods on a new test corpus and a new training corpus is shown in Sections 4.4 and 4.5. Section 4.6 wraps up with an evaluation of the effect of the training corpus size on the results of both test corpora.

4.1 Baseline Results

As a baseline consider the WERs of the OCR output on the unmodified images of the documents from the Eisenhower Communiqués test set and the Enron training set. Each document image in the Eisenhower Communiqués test set, as well as all corpora in this work, was recognized using four OCR engines: Abbyy FineReader 10, OmniPage 18, Adobe Acrobat Pro X, and Tesseract 3. The resulting recognition hypotheses were evaluated using the NIST Sclite tool to compute the number of correctly recognized tokens, as well as substitutions, insertions, and deletions for each document. These results, shown in the first numerical column of Table 4, are a reference point for evaluating the effectiveness of the new techniques and corpora introduced in this paper.

Our previous work showed the improvement in the WER using a trained machine learner with the alignment of the output from multiple OCR engines. Reinterpreting these previous results using micro-averaging techniques, the underlined entries in Table 4 show the decreasing WER as additional OCR outputs are added to the alignment. The order of the OCR output alignment was determined by the order of increasing WER on the Enron training set.

4.2 New Features

In addition to the baseline feature types described in Section 3.3, this paper adds three new feature types for consideration by the machine learner: voting, dictionary lookup, gazetteer lookup, identifying numbers, and the lexical features of the training set. This paper adds three new features:

1. RecurSim, a binary feature indicating whether a token occurs more than once.
2. RecurBucket#, a multivalued feature dividing the number of times a token appears into one of ten buckets, numbered 0 to 9, with each bucket containing approximately the same number of tokens. The higher the bucket number, the fewer times the individual tokens in that bucket appeared in the corpus. (See Section 4.2.1 for more details.)
3. Order, the machine learner is trained using either an order 0 or an order 1 conditional random field (CRF). The order 0 CRF only considers the current column of hypotheses when deciding on the label. The order 1 CRF considers the previous label in addition to the current column features.

4.2.1 Recurring Features

The recurring features (RecurSim and RecurBucket#) are calculated on the training and test sets since the contents of the OCR outputs is available without violating a restriction on using the gold standard transcription of the documents in the corpus. The simple recurring feature (RecurSim) is created for every token that appears more than once anywhere in the corpus. The hope is that by tracking recurring features in the corpus out of vocabulary tokens that do not appear in the dictionary or gazetteer will be captured.

The bucket recurring feature (RecurBucket#) divides up the range of recurring feature counts into ten buckets tagged with the labels RecurBucket0 to RecurBucket9. The method for assigning bucket labels calculates the number of times a given token appears in the OCR of the corpus. For example in the combined OCR of the Eisenhower test corpus there are 3,931 different tokens that each appear twice, one of which is “SCHEUERN” and is likely a valid recognition by the OCR engines of a town by that name in Germany. (See Figure 5.) The buckets are assigned labels in an ascending order such that there are approximately the same number of recurring tokens in each bucket. For the Eisenhower test set the bucket assignments are found in Table 2. The simpler “RecurSim” feature is assigned to all tokens that recur within the corpus, so RecurBucket0 through RecurBucket9 would all be mapped to RecurSim.
4.2.2 CRF Order

Previously, the machine learner used an order 0 conditional random field (CRF), also called a log-linear classifier. This means that only the features of the current column are used to select the label assigned to the features from the hypothesis column. The selected label, which is one of the OCR engines or the label “NONE”, determines which OCR hypothesis to select or whether to select none of the OCR outputs.

For this paper we have added a new set of models, trained using the same training sets but modeled using an order 1 CRF. The model considers not only the features found in the current hypothesis column, but also the label that was selected previously. The results will clearly identify whether the order 0 or the order 1 CRF model is being used.

4.3 Results of the New Features

One of the goals of this research was to determine the contribution of the lexical features learned from the training set to the overall performance of the machine learner. To this end we group features into sets, including and excluding both the RecurSim, RecurBucket#, and the Lexical feature types. The set names and features found in each set are found in Table 3. Refer to Section 3.3 for an explanation of the feature names.

The results across all feature set groupings and CRF orders may be seen in Table 4. To orient the reader, the previously published results are underlined in the “Order 0 CRF” section of the table. Note the italicized entries which indicate improvements within a given OCR alignment over the previous results.

---

‡ The ground truth transcription of the Eisenhower test set consists of 145,346 words in 605 documents. A WER reduction of 0.01% constitutes 15 tokens that are corrected across the corpus, consisting of insertions that are eliminated or words that are corrected.
Table 3. The grouping of features used in various model configurations.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Feature Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Set</td>
<td>Voting, Dictionary, Gazetteer, and Number</td>
</tr>
<tr>
<td>RecurSim Set</td>
<td>All features found in Base Set along with the RecurSim feature type</td>
</tr>
<tr>
<td>RecurBucket# Set</td>
<td>All features found in the Base Set with the RecurBucket# feature type</td>
</tr>
<tr>
<td>Lexical Set</td>
<td>All features found in the Base Set with the Lexical feature type</td>
</tr>
<tr>
<td>Lexical+RecurSim Set</td>
<td>A combination of the Base Set, Lexical Set, and the RecurSim feature type</td>
</tr>
<tr>
<td>Lexical+RecurBucket# Set</td>
<td>A combination of the Base Set, Lexical Set, and the RecurBucket# feature types.</td>
</tr>
</tbody>
</table>

Table 4. Eisenhower test set WERs with the Enron training set including the new features. The underlined entries correspond to the results previously published. Italicized entries indicate improvement over previous results. The bolded entry is the lowest WER in the table.

<table>
<thead>
<tr>
<th>Alignment Order</th>
<th>Base</th>
<th>RecurSim</th>
<th>RecurBucket#</th>
<th>Lexical</th>
<th>Lexical+RecurSim</th>
<th>Lexical+RecurBucket#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order 0 CRF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abbyy + OmniPage</td>
<td>21.91%</td>
<td>22.20%</td>
<td>22.15%</td>
<td>18.31%</td>
<td>18.13%</td>
<td>18.79%</td>
</tr>
<tr>
<td>Abbyy + OmniPage + Tesseract</td>
<td>17.47%</td>
<td>17.15%</td>
<td>17.47%</td>
<td>17.52%</td>
<td>17.29%</td>
<td>17.45%</td>
</tr>
<tr>
<td>Abbyy + OmniPage + Tesseract + Adobe</td>
<td>17.80%</td>
<td>17.55%</td>
<td>17.78%</td>
<td>16.64%</td>
<td>16.23%</td>
<td>16.46%</td>
</tr>
<tr>
<td>Order 1 CRF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abbyy + OmniPage</td>
<td>21.56%</td>
<td>21.76%</td>
<td>21.68%</td>
<td>17.70%</td>
<td>17.61%</td>
<td>17.70%</td>
</tr>
<tr>
<td>Abbyy + OmniPage + Tesseract</td>
<td>17.68%</td>
<td>17.59%</td>
<td>17.70%</td>
<td>17.80%</td>
<td>17.62%</td>
<td>17.75%</td>
</tr>
<tr>
<td>Abbyy + OmniPage + Tesseract + Adobe</td>
<td>17.82%</td>
<td>17.74%</td>
<td>17.79%</td>
<td>17.49%</td>
<td>17.17%</td>
<td>17.42%</td>
</tr>
</tbody>
</table>

Observe that the order 1 CRF is not in general an improvement over the order 0 CRF. With the exception of results using the Abbyy+OmniPage alignment all of the other results in the table have an increased WER. Further, the best result of 17.42% in the order 1 CRF is the Abbyy+OmniPage+Tesseract+Adobe alignment with the Lexical+RecurBucket# Set is significantly higher than the best result in the order 0 CRF table at 16.23%. Based on this, the order 1 CRF will not be included in the results the follow since it is not showing an improvement over the order 0 CRF.

Clearly the Lexical features, as found in the three feature sets Lexical Set, Lexical+RecurSim Set, and Lexical+RecurBucket# Set, show improvement over the feature sets without the Lexical features, yielding a 6.52% relative improvement between the Base Set of the Abbyy+OmniPage+Tesseract+Adobe alignment and the Lexical Set as shown in Table 4. In addition, the recurring features in conjunction with the Lexical features are superior to the Lexical features alone with the Lexical+RecurSim Set having the greatest improvement, showing an additional 2.30% relative improvement over the Abbyy+OmniPage+Tesseract+Adobe alignment and the Lexical Set mentioned above. Overall the Lexical+RecurSim Set performs best on the Eisenhower test set and the Enron training set. We will proceed with the Lexical Set and recurring feature sets as we compare results with the new test corpus, the 19th Century Mormon Article Newspaper Index.

4.4 Results on a Different Test Corpus

The 19th Century Mormon Article Newspaper Index, described in Section 3.1.2, consists of 208,630 words in 1,055 documents digitized to 8-bit grayscale, in contrast to the Eisenhower dataset which is effectively bitonal. The results using the selected feature sets from the previous section are found in Table 5.

\[^{5}\text{A reduction of 0.01% in the WER on the 19th Century Mormon Article Newspaper Index results in 21 tokens that are corrected across the corpus.}\]
Table 5. Results on the Eisenhower and the 19th Century Mormon Article Newspaper Index test sets using a CRF model trained on the Enron, Reuters, and combined training sets. The bold entries are the lowest WERs for each section of the table. The underlined entries are the lowest WERs for their respective test sets.

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Feature Sets</th>
<th>Eisenhower Test Set</th>
<th>19thCMAN Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lexical</td>
<td>Lexical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+RecurSim</td>
<td>+RecurBucket#</td>
</tr>
<tr>
<td><strong>Enron Training Set</strong></td>
<td>Abbyy+OmniPage</td>
<td>18.51%</td>
<td>18.13%</td>
</tr>
<tr>
<td></td>
<td>Abbyy+OmniPage+Tesseract</td>
<td>17.52%</td>
<td>17.29%</td>
</tr>
<tr>
<td></td>
<td>Abbyy+OmniPage+Adobe</td>
<td>16.64%</td>
<td><strong>16.23%</strong></td>
</tr>
<tr>
<td><strong>Reuters-21578 Training Set</strong></td>
<td>Abbyy+OmniPage</td>
<td>19.91%</td>
<td>20.14%</td>
</tr>
<tr>
<td></td>
<td>Abbyy+OmniPage+Tesseract</td>
<td>16.29%</td>
<td>16.12%</td>
</tr>
<tr>
<td></td>
<td>Abbyy+OmniPage+Adobe</td>
<td>16.41%</td>
<td>16.71%</td>
</tr>
<tr>
<td><strong>Combined Training Set</strong></td>
<td>Abbyy+OmniPage</td>
<td>20.01%</td>
<td>20.02%</td>
</tr>
<tr>
<td></td>
<td>Abbyy+OmniPage+Tesseract</td>
<td>16.88%</td>
<td>17.15%</td>
</tr>
<tr>
<td></td>
<td>Abbyy+OmniPage+Adobe</td>
<td>16.63%</td>
<td><strong>16.56%</strong></td>
</tr>
</tbody>
</table>

The monotonic improvement in WER seen on the Eisenhower test set using the Lexical+RecurSim feature set is not reflected in the 19thCMAN test set using the Enron training set. In the 19thCMAN test set, the addition of the Tesseract OCR increases the WER above the Abbyy+OmniPage alignment results across the board for all of the feature sets. Unlike the Eisenhower test set, 19thCMAN appears to have a sensitivity to the relatively high WER of the Tesseract OCR (56.08%). Adding the Adobe OCR output improves the resulting WER to a level equal to or below both the Abbyy FineReader WER and the Abbyy+OmniPage alignment, even though the WER of Adobe on the 19thCMAN test set (68.95%) is higher than that of Tesseract. The conclusion here is that although the high WER OCR of Tesseract and Adobe in the training set were able to contribute to lowering the WER for the Eisenhower test set, in combination they did not contribute in the same way with the 19thCMAN treat set. A possible solution may be to eliminate from the training set documents with high WERs. Since the training set includes alignments of documents from multiple OCR engines, this may decrease the size of the training set since if a document is eliminated due to a high WER from one OCR engine, it would need to be eliminated from all of the OCR engine contributions to the training set. Section 4.6 explores whether the full contents of the training set are needed to maintain the level of performance seen so far.

### 4.5 Results on New Training Corpora

New in this paper, the Reuters-21578 training set described in Section 3.1.3 is a synthetic dataset consisting of grayscale images, which is in contrast to the Enron training set which had previously been binarized. The hope was that since the 19thCMAN test set was grayscale, that the Reuters-21578 training set would contribute to improving the over WER. Note that the effects of the high WER OCR outputs from Tesseract and Adobe Pro X seem more pronounced with this training set. The WER results for Abbyy+OmniPage were the best and further adding Tesseract and Adobe Pro X to the alignment each increased the WER. Overall, however, the Reuters-21578 training set showed better results than the Enron training set.
The last rows of Table 5 show the results of merging both the Enron and the Reuters-21578 training sets. As more training data is available, there is no improvement on the Eisenhower test set and a small improvement of only 0.01% for the 19thCMAN test set. The conclusion is that there is a point where more training set vectors does not necessarily improve the outcome. Overall the best results were seen with the Reuters-21578 training set, although not consistently with the complete set of OCR alignments.

4.6 Results of Sweeping the Size of the Training Sets

Exploring the observation from the last section, that the increased size of the training set does not necessarily improve the WER outcome, we sweep the size of the Reuters-21578 training set from 0.01% to 100% to explore how the WER of the Eisenhower and 19thCMAN test sets varies as the training set increases in size. We selected the best result on the Eisenhower test set across all training sets, which was the Abbyy+OmniPage+Tesseract alignment using the Lexical+RecurBucket# Set as seen in Table 5. We selected ten proportion values (0.01%, 0.1%, 0.2%, 0.5%, 1%, 2%, 5%, 10%, 20%, and 50%) of the training set, which has 585,291 feature vectors. At each proportion value we took five random samples from the full training set to create 50 new training sets from which order 0 CRF models were created. The results shown in Figures 6 and 7 are the average WER, as well as the minimum and maximum WERs for the five models of each proportion value. When selecting a proportion size of only 0.01% of the total training set, only 61 feature vectors on average are included in the models created.

Of interest on the Eisenhower results is that beginning with a training set proportion size of only 2.0% the average WER of the resulting error corrected output is less than the WER using the entire test corpus. Given that the full test set takes a significant amount of time to train, the five 2.0% test sets are considerably faster to train. Further the models created from the 2.0% test sets consist of on the order of 11,500 features while the full model consists of over 217,000 features. Clearly the complexity of the full model does not necessarily reward us with better results.

Regarding the 19thCMAN test set, similar to the Eisenhower test set, the full effect of the Reuters-21578 training set is visible between 1% and 10% of the total training set. Interestingly, superior results are possible given individual training set proportions within the same range.

5. CONCLUSIONS

In general the results seen in previous work are also seen with the new test set and training set. Although the methodologies used in previous papers still produced good results they do not work consistently across the board. The 19thCMAN test
set seemed sensitive to the high WER OCR engines included in the training set, but we demonstrated that the size of the training set can be reduced, potentially eliminating the troublesome high WER documents, potentially improving the end results. Future work will include error analysis of how the models of the Enron and Reuters-21578 training sets differ in their performance on the Eisenhower and 19thCMAN test sets.

Acknowledgments

The authors express their appreciation to the Fulton Supercomputing Lab and the Harold B. Lee Library, L. Tom Perry Special Collections, at Brigham Young University, without whose support this research would not have been possible. We thank the employees of the Digital Initiatives Lab of the Harold B. Lee Library for their extensive work in digitizing, OCRing, and correcting the OCR text of the Nineteenth Century Mormon Article Newspaper Index. Lastly, we express gratitude to Patricia A. Fraade of the Harold B. Lee Library for her extensive work in assembling and curating the Nineteenth Century Mormon Article Newspaper Index and Dr. Richard Hacken for curating the Eisenhower Communiqués.

REFERENCES


