2011 International Conference on Document Analysis and Recognition

Progressive Alignment and Discriminative Error Correction for Multiple OCR Engines

William B. Lund
Computer Science Department
and the Harold B. Lee Library
Brigham Young University
Provo, Utah 84602, USA
bill_lund@byu.edu

Daniel D. Walker
Computer Science Department
Brigham Young University
Provo, Utah 84602, USA
danl4@cs.byu.edu

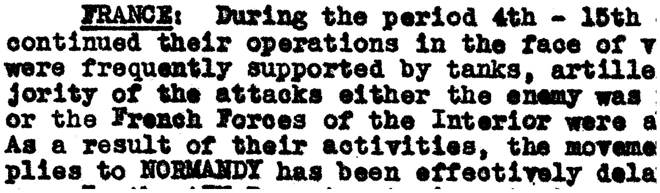
Eric K. Ringger
Computer Science Department
Brigham Young University
Provo, Utah 84602, USA
ringger@cs.byu.edu

Abstract—**This paper presents a novel method for improving optical character recognition (OCR). The method employs the progressive alignment of hypotheses from multiple OCR engines followed by final hypothesis selection using maximum entropy classification methods. The maximum entropy models are trained on a synthetic calibration data set. Although progressive alignment is not guaranteed to be optimal, the results are nonetheless strong. The synthetic data set used to train or calibrate the selection models is chosen without regard to the test data set; hence, we refer to it as "out of domain." It is synthetic in the sense that document images have been generated from the original digital text and degraded using realistic error models. Along with the true transcripts and OCR hypotheses, the calibration data contains sufficient information to produce good models of how to select the best OCR hypothesis and thus correct mistaken OCR hypotheses. Maximum entropy methods leverage that information using carefully chosen feature functions to choose the best possible correction. Our method shows a 24.6% relative improvement over the word error rate (WER) of the best performing of the five OCR engines employed in this work. Relative to the average WER of all five OCR engines, our method yields a 69.1% relative reduction in the error rate. Furthermore, 52.2% of the documents achieve a new low WER.**

Keywords-**Optical character recognition software; Error correction; Machine learning; Multiple sequence alignment; Progressive text alignment; Synthetic training data set**

I. Introduction

In pursuit of high quality digital versions of historical documents, this paper demonstrates the extent to which improvements in the recognized (digital) text are possible as additional OCR hypotheses are incorporated from multiple engines through progressive alignment (cf., [15], [16]). This paper is organized as follows. Section II discusses related work. Methods used for alignment and the baseline results are in Section III. Section IV discusses the creation of an out-of-domain synthetic data set used to train a maximum entropy model for error correction. Our conclusions are presented in Section V.



FRANCE: During the period 4th - 15th
continued their operations in the face of v
were frequently supported by tanks, artille
jority of the attacks either the enemy was
or the French Forces of the Interior were a
As a result of their activities, the moveme
plies to NORMANDY has been effectively dela

Figure 1. From "Periodical 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%.

II. Related Work

There is a significant body of published work on the use of multiple inputs for OCR error correction. Klein and Kopel [9] note that OCR engines show wide variation in the types of errors made and that voting between engines is effective in identifying accurate OCR word hypotheses. Voting among multiple hypotheses has also been explored by Lopresti and Zhou [11], in which multiple scans of the same document were evaluated by the same OCR engine and voting was employed to make the final selection. Lin [10] uses multiple OCR engines to recognize the same document, aligning the OCR text output, with majority voting on the output. Our previous work introduced an efficient exact alignment algorithm [12] and domain-specific training [13] to correct OCR using three engines (two commercial and one open source). Boschetti et al. [5] also align multiple OCR outputs, selecting characters using a naïve Bayes classifier.

In the domain of genetic multiple sequence alignment problems, progressive alignment has been shown to be highly effective in achieving good, although not guaranteed optimal results. (See Moretti et al. [15] and Notredame (2007) [16].) Spencer and Howe [18] apply progressive alignment to textual variants of ancient and historical documents while Feng and Manmatha [7] use a Hidden Markov Model to align the OCR of a full book-length text to an existing electronic version.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.303

A contribution of this paper is the use of supervised, discriminative machine learning methods to choose among all hypotheses. Maximum entropy models have been used previously to select among multiple parses returned by a generative model (e.g. [6]). In this work the models are learned on a synthetic, out-of-domain calibration data set, created and computationally degraded according to the methods proposed by Sarkar, Baird, and Zhang [17] and Baird [3].

III. Methods and Results on the Eisenhower Communiqués

A. Data

The historical documents used in this paper are the Eisenhower Communiqués [8], a collection of 610 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 1 for an example.) A manual transcription of these documents serves as the gold standard for evaluating the word error rates of the OCR. Two-thirds of the documents have been assigned randomly to an evaluation set for this research.[1] One-third of the documents are reserved for future research. We employ no Eisenhower Communiqués data as training data in this work and instead focus on the scenario of recovering document text in the absence of in-domain training data, using an out-of-domain synthetic calibration set.

B. Baseline OCR

Each of the document images in the Eisenhower Communiqués evaluation set was recognized using five OCR engines: Abbyy FineReader for Windows (version 10), OmniPage Pro X for Mac OS X, Adobe Acrobat Pro for Mac OS X (version 9), ReadIris Pro for Mac OS X (version 11.6), and Tesseract (version 1.03), an open source OCR system. The resulting recognition hypotheses were evaluated using the NIST Sclite [1] tool to compute word error rates (WER) and lattice word error rates (LWER). The baseline WERs for the Eisenhower Communiqués data set can be found in the top half of Table I. These baseline results are a reference point for evaluating the effectiveness of the techniques introduced in this paper. The expectation is that the types of errors as well as the types of successful recognition will vary across engines. Leveraging these variations to correct recognition errors is the goal of this research.

	Word Error Rates				
Eisenhower Communiqués	Abbyy	OmniPage	Adobe	ReadIris	Tesseract
Mean	18.24%	30.02%	51.78%	54.64%	67.78%
	Average WER across all OCR engines: 44.49%				
Minimum	1.87%	1.45%	2.38%	2.38 %	2.01%
Maximum	84.71%	112.68%	151.22%	206.75%	1017.11%
Enron Synthetic Calibration Set	Abbyy	OmniPage	Adobe	ReadIris	Tesseract
Mean	25.02%	31.92%	67.57%	69.62%	56.03%
	Average WER across all OCR engines: 50.03%				
Minimum	0.34%	1.34%	6.02%	5.42 %	4.19%
Maximum	166.34%	205.94%	170.79%	200.00%	176.73%

Table I. Baseline word error rates for the OCR engines on all documents in the evaluation dataset of the Eisenhower Communiqués and the Enron synthetic calibration data set. Note that WERs of greater than 100% are possible due to multiple insertions not found in the reference text.

C. 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 [16] and textual variance analysis [18]. 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 and Howe [18] 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.

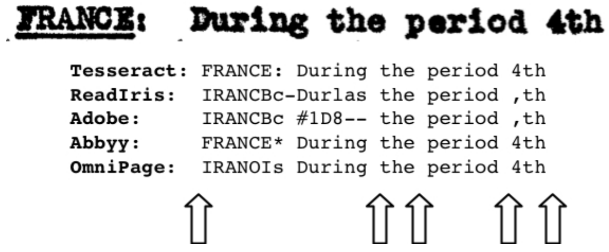
In this paper, the order of the alignment, unless indicated otherwise, is a greedy approximation of the guide tree based on sequence similarity of the calibration set (discussed in Section IV-A); specifically: Abbyy FineReader and OmniPage Pro X, then Adobe Acrobat Pro and ReadIris Pro, and lastly Tesseract. However, in order to show the effect of adding OCR engines individually, the individual OCR hypotheses were introduced one at a time (progressively) to the overall alignment in a manner consistent with the guide tree.

D. Lattice Word Error Rates

From the final, overall alignment of the five OCR outputs, we create columns of hypotheses delimited by consensus on white space. Our intention is that each column captures aligned words from the document image. (See Figure 2 for an example.) Each aligned column is a list of hypotheses from which to select a single best hypothesis.

The Lattice WER (LWER) is an oracle calculation: for each column, if any of the OCR hypotheses in the column matches the truth in the transcript, it is considered a correct match. As more good hypotheses are added to the lattice, the LWER is reduced, as shown in Table II and in Figure 3. There are several interesting points to observe in the LWER

[1] Previous papers [12], [13] using the Eisenhower data set divided the current evaluation set into a training set and a development test set, which accounts for differences in reported development test set WERs.



```
Tesseract: FRANCE: During the period 4th
ReadIris:  IRANCBc-Durlas the period ,th
Adobe:     IRANCBc #1D8-- the period ,th
Abbyy:     FRANCE* During the period 4th
OmniPage:  IRANOIs During the period 4th
```

Figure 2. From “Periodical Communiqué No. 1”, an example of aligned sequence hypotheses from five OCR engines. The arrows indicate points of agreement on white space among the aligned sequences. The text between an adjacent pair of arrows constitutes an aligned column of hypotheses. The hyphen “-” character in a sequence represents a gap aligned with characters in the other sequences.

Order by	Number of Aligned OCR Sequences				
OCR WER	1	2	3	4	5
Low to High	18.24%	12.01%	11.28%	11.14%	9.27%
High to Low	67.78%	27.10%	24.08%	16.44 %	9.58%

Table II. Improvements to the Lattice WER for the Eisenhower Communiqués as OCR outputs are progressively added. Row 1: from best to worst. Row 2: from worst to best.

results in row 1 of the table. First, even though Tesseract has the highest overall WER, it still has information to contribute resulting in a decrease in the LWER from 11.14% for a 4-way alignment of Abbyy, OmniPage, Adobe, and ReadIris to 9.27% when Tesseract is added for a 5-way alignment. This result indicates that even higher error rate information sources can potentially contribute to reducing the overall error rate. Second, the order in which the OCR outputs are added to the alignment does little to affect the ultimate Lattice WER when all five engines are included. The small difference between the two ultimate outcomes (shown in the last column of Table II and the 5-way alignment in Figure 3) can be attributed to the sub-optimalities of the progressive alignment method. The Lattice WER is a lower bound on what is possible given the information contained in the joint alignment of the multiple OCR outputs. Selecting the correct hypotheses within the aligned columns is the remaining task.

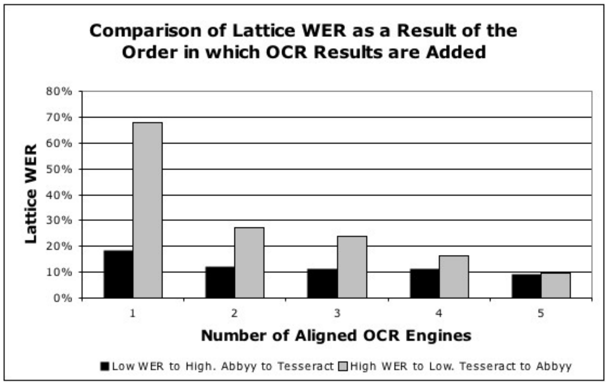


Figure 3. Improvements to Lattice WER for the Eisenhower Communiqués

IV. Machine Learning Methods and Results with Out-of-Domain Training

A. *Enron Synthetically Generated Data Set*

Our goal in this research is to explore how well OCR errors can be corrected without requiring a domain-specific training set since such data may be unavailable or too expensive to acquire. In order to use modern discriminative supervised machine learning methods, we used an out-of-domain calibration set. For the calibration set, we created a synthetic data set from the 2001 Topic Annotated Enron Email Data Set [4], a corpus available from the Linguistic Data Consortium (LDC). The choice of the Enron data was essentially arbitrary and reflects our commitment to having a trained model very unfamiliar with our evaluation data. From the digital text of each document, a TIFF document image was generated and randomly degraded using techniques inspired by Baird [3] and Sarkar et al. [17]. Each image was produced in the following way:

1) Create an image from the text as a bi-tonal document at 1500 dots per inch (dpi), which is five times the target resolution of 300 dpi.
2) Introduce spatial sampling error by translating the entire image between one to five pixels in both the x and y axes, which introduces randomness when subsampling.
3) Blur the image using a Gaussian convolution kernel in which the value for each pixel is taken to be the weighted average of its neighboring pixels.
4) Subsample the document to 300 dpi.
5) Simulate pixel sensor sensitivity by adding a value for each pixel individually drawn from a Gaussian distribution with a mean of zero and a standard deviation of 0.025.
6) Choose a random threshold from a truncated normal distribution (between 0.1 and 0.4) with mean 0.225 and a standard deviation of 0.11418, and binarize the document using this threshold.

Note that documents were synthesized with particular parameter ranges chosen independently to reflect the kinds of noise we expected to see in our data, allowing for wide variations. The WERs on the Enron calibration set were comparable to the rates on Eisenhower Communiqués (See Table I) ranging from a mean of 25.02% (Abbyy) to a mean of 69.92% (ReadIris).

B. *Training Maximum Entropy Model Using the Enron Data Set*

We employ modern supervised discriminative machine learning methods trained on the calibration 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 Enron calibration set 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. We extracted the following kinds of features 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, and
- *Spell Checker*: an additional hypothesis generated by Aspell [2] 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 OCR engine with the lowest WER from the calibration set. Consider the following example column from the Enron calibration set: {Abbyy: "Precipitation", OmniPage: "Precipitation", Adobe: "Prcdpitalion", ReadIris: "Prccipitalion", Tesseract: "Precipitation:"}. The Spell Checker also provided the hypothesis: "precipitation". The following features are extracted from this column:

Type of Feature	Feature Values
Word Hypotheses:	T:Precipitation: R:Prccipitalion D:Prcdpitalion A:Precipitation O:Precipitation S:precipitation
Voting	VoteAOS VoteAO VoteAS VoteOS
Dictionary:	DictT DictA DictO
Spell Checker:	Spell
Training Label:	A

Note that "DictA" (and so forth) indicates that the entry from each respective OCR engine is found in the dictionary. Leading and trailing punctuation is removed from the hypothesis before checking in the dictionary. To produce a "Voting" feature the match must be exact, including punctuation. During training the label assigned to these features is "A", meaning that Abbyy's is the correct hypothesis to be selected. Once all of the feature vectors have been extracted, we use the maximum entropy learner in the Mallet [14] toolkit to train a maximum entropy (a.k.a., multinomial logistic regression) model to predict choices on unseen alignment columns.

	Word Error Rates				
Method	Abbyy Alone	+OmniPage	+Adobe	+ReadIris	+Tesseract
Machine Learning	18.24%	16.54%	15.09%	14.59%	13.76%

Table III. WER from the Eisenhower evaluation set for a trained machine learning method to select hypotheses from the aligned lattices as additional OCR outputs are added.

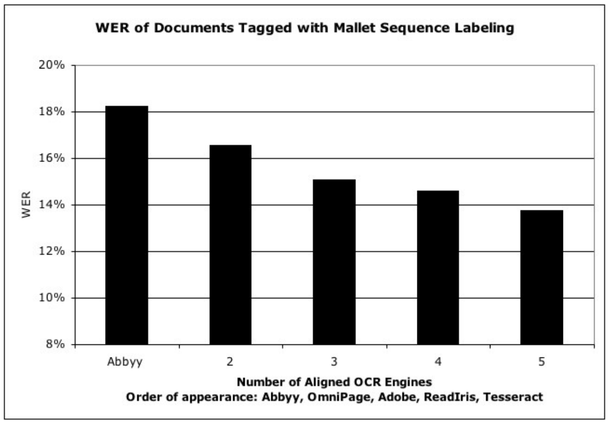


Figure 4. Decreasing mean WER on the Eisenhower evaluation set using machine learning methods and an out-of-domain training set. (See the last row of Table III.)

C. Machine Learning System Results

The following process is depicted in Algorithm 1. Using the model created with the Enron calibration set, our algorithm assigns a label to each column of hypotheses in each document in the Eisenhower evaluation set. The maximum entropy learner in Mallet indicates which OCR hypothesis to select from the column. The selected hypotheses are then assembled for each document as the corrected output and evaluated by Sclite. It should be noted that Sclite does not distinguish between upper and lower case characters. The function $features()$ in Algorighm 1 returns the features of the aligned column, c, of hypotheses from the OCR engines and $argmax^*()$ performs the expected function while breaking ties as described in Section IV-B. The results can be seen in Table III and in Figure 4. From Table III the lowest WER achieved was 13.76%, which is a 69.1% relative reduction from 44.49%, the mean WER of all OCR engines, and a 24.6% relative reduction from 18.24%, the WER of the best OCR engine (see Table I). Also observe that with each addition of an OCR output, the mean WER on the Eisenhower evaluation set decreases.

Another indication of the ability of the system to take advantage of information provided as new OCR outputs are added is the number of documents that have a reduced WER

Algorithm 1 recognizeMultipleOCR(d, m, E)

```
INPUT: document: $d$
       model: $m$
       OCR engines set: $E$
$Alignment_d \leftarrow progressiveAlign(d, E)$
$Columns_d \leftarrow splitOnWhitespace(Alignment_d)$
$Transcription_d \leftarrow nil$
for all $c \in Columns_d$ do
    // $c = \{h_a, h_b, h_c, \ldots, h_n\}$
    // $select(c, m) = argmax^*_{h_i \in c} P_m(h_i|features(c))$
    $selection \leftarrow select(c, m)$
    $Transcription_d \leftarrow append(Transcription_d, selection)$
end for
return $Transcription_d$
```

Percentage of Documents Reducing WER				
	+OmniPage	+Adobe	+ReadIris	+Tesseract
Reduction from previous alignment	44.16%	83.38%	59.22%	65.19%
New low WER of all previous alignments	44.16%	50.91%	39.22%	52.21 %

Table IV. Beginning with Abbyy FineReader, as OCR outputs are progressively added to the aligned sequences, at each step there is a significant percentage of documents that reduce their WER either from the previous alignment or as a new overall low WER.

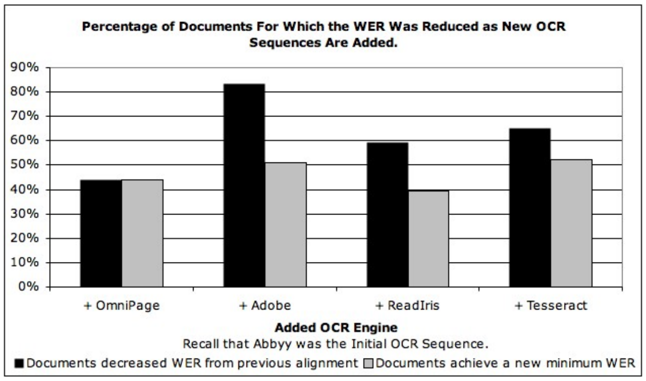


Figure 5. As OCR outputs are added to the already aligned sequences, at each step the percentage of documents that reduce their WER, alongside the WER of that OCR engine. (See Table IV).

at each progressive step. This is calculated in two ways: first, the percentage of documents that have a lower WER than in the previous step in the progressive alignment, and second, the percentage of documents that achieve a new overall lower WER at that step. Table IV and Figure 5 show these results. Note that even after having aligned Abbyy, OmniPage, Adobe, and ReadIris, still 52.21% of the documents have a lower minimum WER with the addition of the Tesseract OCR output, despite Tesseract's 67.78% WER on the Eisenhower evaluation set. We conclude that OCR outputs with even very high WERs can significantly contribute to WER reduction.

V. Conclusions

We have documented the degree to which information from multiple OCR engines can be used in an aligned lattice of OCR hypotheses to improve OCR performance. As more OCR engines are included in the alignment, the Lattice WER decreases, even when adding OCR outputs with significant WERs. Thus, progressive alignment provides a usable alternative to exact alignment for processing five OCR sequences. Ultimately, when incorporating multiple OCR engines, the order in which they are added makes little difference on the Lattice WER in the final outcome. Machine learning techniques succeed in leveraging the available information in the lattice: using out-of-domain training data is effective for training a maximum entropy model to select correct hypotheses from the aligned OCR sequences. This research made use of an innovative means for creating a domain-independent calibration training set, which was shown to be successful when used to build models for use with the Eisenhower Communiqués, a historical data set with significant degradation. This work presents a compelling new method for producing digital text from historical documents.

References

[1] J. Ajot, J. Fiscus, N. Radde, and C. Laprun. asclite–Multi-dimensional alignment program. http://www.nist.gov/speech/tools/asclite.html, 2008.

[2] K. Atkinson. GNU Aspell. http://aspell.net/, 2008.

[3] H. Baird. The State of the Art of Document Image Degradation Modelling. In *Digital Document Processing*, pages 261–279. Springer, 2007.

[4] M. W. Berry, M. Browne, and B. Signer. 2001 Topic Annotated Enron Email Data Set. http://www.ldc.upenn.edu/, June 2007.

[5] F. Boschetti, M. Romanello, A. Babeu, D. Bamman, and G. Crane. Improving OCR accuracy for classical critical editions. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, Corfu, Greece, Sept. 2009. Springer Verlag.

[6] E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 173–180, Ann Arbor, MI, June 2005.

[7] S. Feng and R. Manmatha. A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. pages 109–118, Chapel Hill, NC, June 2006.

[8] D. R. Jordan. Daily battle communiques, 1944-1945. Harold B. Lee Library, L. Tom Perry Special Collections, MSS 2766, 1945.

[9] S. T. Klein and M. Kopel. A Voting System for Automatic OCR Correction. In *Proceedings of the SIGIR 2002 Workshop on Information Retrieval and OCR*, Aug. 2002.

[10] X. Lin. Reliable OCR Solution for Digital Content Re-mastering. San Jose, CA, Jan. 2002.

[11] D. Lopresti and J. Zhou. Using consensus sequence voting to correct OCR error. *Computer Vision and Image Understanding*, 67(1):39–47, 1997.

[12] W. B. Lund and E. K. Ringger. Improving Optical Character Recognition through Efficient Multiple System Alignment. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 231–240, Austin, TX, USA, 2009. ACM.

[13] W. B. Lund and E. K. Ringger. Error Correction with In-Domain Training Across Multiple OCR System Outputs. In *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011)*, Beijing, China, Sept. 2011.

[14] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

[15] S. Moretti, F. Armougom, I. M. Wallace, D. G. Higgins, C. V. Jongeneel, and C. Notredame. The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. *Nucl. Acids Res.*, 35(suppl_2):W645–648, July 2007.

[16] C. Notredame. Recent Evolutions of Multiple Sequence Alignment Algorithms. *PLoS Computational Biology*, 3(8):e123, 2007.

[17] P. Sarkar, H. S. Baird, and X. Zhang. Training on Severely Degraded Text-Line Images. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 1*, page 38. IEEE Computer Society, 2003.

[18] M. Spencer and C. Howe. Article: Collating Texts Using Progressive Multiple Alignment. *Computers and the Humanities*, 38(3):253–270, Aug. 2004.

[19] L. Yost. U.S. Board on Geographic Names. http://geonames.usgs.gov/, May 2009.