Error Correction with In-Domain Training Across Multiple OCR System Outputs

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

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

***Abstract*—Optical character recognition (OCR) systems differ in the types of errors they make, particularly in recognizing characters from degraded or poor quality documents. The problem is how to correct these OCR errors, which is the first step toward more effective use of the documents in digital libraries. This paper demonstrates the degree to which the word error rate (WER) can be reduced using a decision list on a combination of textual features across the aligned output of multiple OCR engines where in-domain training data is available. This research was performed on a data set for which the mean WER across the three OCR engines employed is 33.5%, and the lattice word error rate is 13.0%. Our correction method leads to a 52.2% relative decrease in the mean WER and a 19.5% relative improvement over the best single OCR engine, as well as an improvement over our previous work. Further, our method yields instances where the document WER approaches and for five documents matches the lattice word error rate, which is a theoretical lower bound given the evidence found in the OCR.**

***Keywords*-Optical character recognition; OCR error correction; Multiple OCR engines; Decision lists**

I. INTRODUCTION

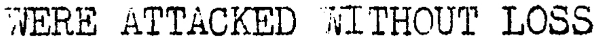
Major digitizing efforts are making pre-digital materials available on-line at an unprecedented scale. Our research leverages the variation among OCR engines (see Figure 1) and additional features of the OCR hypotheses to improve the output beyond what any single OCR engine is capable of. In this case, where in-domain training data is available, we improve upon our previous work [8] and show how a decision list trained on in-domain data using feature combinations reduces the word error rate beyond what is achieved using consensus voting or dictionary matching alone. Further, we explore using a spell checker to suggest additional words for hypotheses that do not appear in the dictionary or gazetteers..

The remainder of this paper is organized as follows. Section II gives background on the problem, information about the data set used, and related work. Section III describes our methodology and presents the results in reducing the document OCR error rate using features alone and in combination. And Section IV summarizes our conclusions and proposes future work.

II. APPROACH

A. Background

The documents used in this paper are the Eisenhower Communiqués [5], a collection of 610 facsimiles of typewritten documents created by the Supreme Headquarters of the 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 very poor. Many documents have artifacts of the duplication process, further complicating the text recognition task. A manual transcription of these documents serves as the gold standard for evaluating the word error rates of the OCR and the feature weighting process described in Section III. The documents have been randomly divided into three sets, each roughly one third of the total collection: a training set, a development test set and a blind test set. The blind test set of documents is reserved for future work.



OCR A: `7JERE ATTACKED WITHOUT LOSS`
OCR B: `WERE  ATTACKED ;ITHouT LOSS`

Figure 1. Poor quality text from Eisenhower Communiqué No. 233a along with OCR output.

B. Related Work

OCR, particularly for historical documents continues to be an area of open research. Hull [4] notes that even small character error rates result in significantly higher word error rates, which negatively affects the usefulness of the OCR in document searching and other tasks. He calculates that a 1.4% character error rate can lead to a 7.0% word error rate in a document with 2,500 characters and 500 words. Kae and Learned-Miller [6] confirm that although the OCR of modern clean documents is effectively a solved problem, older degraded documents present difficulties.

Many approaches have been taken to reduce the error rate of OCR output. Lopresti et al. [7] use consensus voting of characters between multiple scans of the same document recognized by the same OCR software. The National Library of Medicine [16] selected Prime OCR, a commercial system

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.138

that votes on responses from multiple OCR engines, to improve word recognition. The authors' previous work [8] also explored voting in conjunction with feature engineering. The approach reported in this work uses distinct OCR engines to find variations in the text recognized by the systems, and similar to Lopresti, we find that voting plays an important role in making good corrections. Using character confusion learned from the document itself, Kae and Learned-Miller [6] use a weighted English lexicon to provide word hypotheses for unknown tokens. Likewise Strohmaier et al. [15] use a dictionary and a Levenshtein edit distance for post-OCR error correction. Wick et al. [17] explore using a weighted lexicon created from a topic model rather than a dictionary. This research will show that the dictionary matching process in conjunction with aligned hypotheses from multiple engines does in fact result in a lower word error rate, but that using additional features beyond these two improves the results.

Similar to the text error correction task, error correction in speech recognition has employed related methods. For instance, Ringger [12], [13] explored statistical methods in speech recognition post-processing to correct errors. Likewise Mangu et al. [10] took a lattice of alternatives from a speech recognizer and proposed a method for creating a probabilistically consistent lattice for word error minimization. Our work likewise uses a lattice of alternatives, but rather than consisting of alternatives generated by a single source, we use the alignment of multiple sources to populate the lattice.

Combinations of multiple models or systems have been shown to provide improved results. For example, Fiscus [3] votes among multiple speech recognition systems and Nakano et al. [11] use multiple OCR inputs and alignment of lines in the text to improve OCR recognition. We will show that the combination of feature weighting with multiple OCR alignment results in improvements beyond voting or dictionary matching alone.

III. Methodology and Results

A. Baseline OCR Results

Each of the documents in this study was evaluated using three OCR engines, two commercial and one open source, referred to here as OCR A, OCR B, and OCR C respectively. The same digitized image file was evaluated by each OCR engine, and their respective word error rates were calculated with Sclite [1], a tool provided by NIST for use in speech recognition research. Sclite calculated the lattice word error rates and word error rates, shown in Table I.

	OCR A	OCR B	OCR C	Mean	Lattice
	WER	WER	WER	WER	WER
Mean	19.9%	30.4%	50.1%	33.5%	13.0%
Minimum	2.3%	1.5%	3.9%	3.4%	0.7%
Maximum	80.7%	111.7%	1,666.4%	602.2%	78.4%

Table I
Baseline error rates.

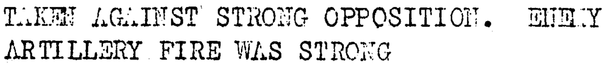
The texts from the three OCR engines are character aligned using the A* algorithm with the Reverse Dijkstra admissible heuristic described by Lund and Ringger [8]. From this character level alignment we construct a lattice of word hypotheses such that wherever there is agreement across all engines on the location of white space we construct a column of hypotheses. The order of hypotheses in the column is determined by the overall accuracy of the OCR engine, with the lowest WER being first. (See Figure 2.) From this lattice of word hypotheses we calculate the lattice word error rates shown in Table I by comparing the true transcription with all of the aligned hypotheses from the OCR output. If any of the hypotheses in the column match the true transcription it is considered a match. This score provides a lower bound on the error rate that is possible using the evidence found only in the OCR outputs.

B. Features of the Aligned OCR Text

The alignment creates columns of alternative hypotheses suggested by the OCR engines. Each of the hypotheses is evaluated for features that may be indications of its accuracy. Following are the features used in the experiments reported in this paper.

1) **Voting** [V:#]: The count of hypotheses within a column that match the current hypothesis exactly. For example, the feature V:3 indicates that the hypothesis in question matches two other hypotheses in the column.
2) **Dictionary** [D]: A binary indicator for whether a hypothesis appears in the Unix dictionary.
3) **Gazetteer** [G]: A binary indicator for whether a hypothesis appears in a gazetteer of European place names.
4) **Number** [N]: A binary indicator for whether a hypothesis is a number.
5) **Recurring** [R]: A binary indicator for whether the hypothesis appears in a list of tokens that do not appear in the dictionary or gazetteer but repeat in the corpus.
6) **Spell** [S:#]: For hypotheses that do not appear in the dictionary or gazetteer, we use the GNU Aspell [2] spell checking software to add additional hypotheses, which may suggest correct words not found in the OCR.

Section III-C below explores the word error rate results when using a single feature to select a hypothesis from the column of alternatives, and Section III-D presents the results when using a decision list trained on a combination of features.



TAKEN AGAINST STRONG OPPOSITION. ENEMY
ARTILLERY FIRE WAS STRONG

A:	TAKE!	[D]	AGAINST	[D]	STRONG	[V:3, D]
B:	T:JcJT	[]	L.G_Il'TST'	[]	STRONG	[V:3, D]
C:	T.;I()?2uI	[]	AGAINST‘	[D]	STRONG	[V:3, D]
S:						

A:	OPPOSITION.	[V:2, D]	31710	[N]
B:	OPPOSiTIOV.	[]	E1ii:Y	[]
C:	OPPOSiTION.	[V:2, D]	EUEJQY	[]
S:	OPPOSITION	[S:1]	ENJOY	[S:1]

Figure 2. Aligned output of the OCR engines from part of Communiqué No. 204 with assigned features. "S:" indicates word hypotheses added by the spell checker.

C. Untrained Single Feature Success Rates

The first question we address is what is the potential for individual features for identifying the correct hypothesis without the aid of parameter training? The features of each hypothesis are calculated, and each hypothesis itself is compared to the true word from the transcription. The results reflecting the potential usefulness of individual features are shown in the columns of Table III at the bottom.

Based on our observations and those of related work, we expected that **Voting** would be a powerful indicator of success in identifying the actual word and the results bear this out. **Dictionary** matches were also a strong indicator of the underlying word from the document, with **Gazetteer** matches having fewer instances but still showing promise in indicating the true word from the document. The result using the dictionary is consistent with our earlier paper [8], which showed a reduced WER when using that feature alone.

D. Combining Features

Encouraged by the potential of our chosen features, we next consider the effects of combining features. Consider instances where a **Vote**:# feature is present (meaning either Vote:3 or Vote:2 is present), but the hypothesis is not found in the dictionary. In Table II we see that when the **Vote**:# feature is false, combined with the **Dictionary** feature, the potential success rate is significantly lower. Likewise, the **Dictionary** feature combined with the absence of a **Vote**:# feature has a lower success rate. Finally, if we combine just the **Vote**:# features and the **Dictionary** feature the combined success rate is higher than those features alone. These potential success rates lead us to believe that features taken together provide a better indicator of whether a hypothesis is more likely to be the true underlying word from the document than features considered individually.

Features	**Instances**	**% Correct**
Vote:# and not **Dictionary**	16,2123	71.4%
Dictionary and not **Vote:#**	6,873	41.3%
Vote:# and **Dictionary**	84,640	96.2%

Table II
SUCCESS RATES OF DISJOINT AND COMBINED FEATURES ON THE TRAINING SET.

Using the in-domain training set, we evaluated the success rates of combinations of features, as shown in Table III. The methodology was similar to that described in Section III-C for individual features; however, in this case we calculate the success rate of combinations of features observed in the training data.

Looking at the results in Table III, note that the combination of the **Vote:3** feature (all three OCR engines output the same hypothesis) combined with the **Dictionary** feature (the hypothesis is found in the dictionary) gives a very high potential success rate of 97.2%, higher than **Vote**:# features or the **Dictionary** feature individually. Of greater interest is the fact that the **Vote:2** feature (two of the three OCR engines agree) combined with the **Dictionary** feature still has a high success rate (93.6%), which is also higher than a **Vote**:# or the **Dictionary** features individually. Overall the **Vote**:# feature for all values, when combined with other features, is a strong indicator of the underlying word from the document. **Vote:3** and **Vote:2** alone, without any other features being present, are significantly less indicative of the underlying word.

Another interesting point from Table III is contra-indicators, specifically feature combinations that would seem to indicate that the hypothesis should be excluded all together. For example, the lack of any features is a strong indicator that the hypothesis is not found in the document.

E. Decision List Training

One way to take advantage of the information learned in Table III is to employ the table as a decision list (defined by Rivest [14]) to score each hypothesis in each column, creating a two step process: 1) Using the training set, learn the correctness percentage for each combination of features that appears in the training set. 2) Apply these learned rates to the development test set as follows: (a) for each word hypothesis in each document in the development test set, calculate the feature set and look-up the correctness score learned from the training set. (b) From the scores of the hypotheses in an aligned column, select the hypothesis with the highest scoring combination of features. (See Figure 4.) To break ties, choose the first hypothesis in the column.

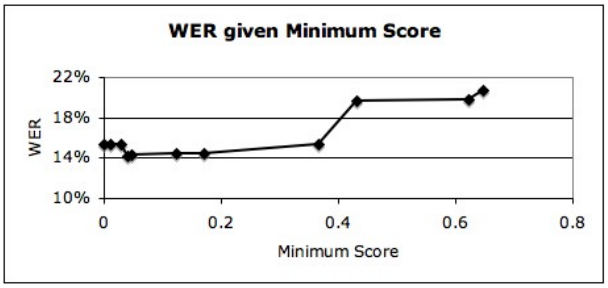
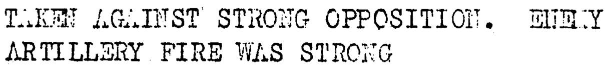


Figure 3. WER on the training set for minimum scores required to be included in the output text.

Is it possible that some features should be an indicator that the hypothesis should be discarded? Not all of the feature combinations have strong potential success rates, and hence doubtful that they are indicative of a true word. Using the combined feature accuracy rates shown in Table III, a naïve approach is to reject all hypotheses with features whose combinations have a score below 50%. On the training set we explored varying the lower bound at which a hypothesis may be selected for output. The lowest WER was achieved when the minimum threshold was set at 4.1%, which excluded hypotheses with **No Features**, **Spell:1** and

	Feature Combinations										Training Set Results	
	Vote:3	Vote:2	Dictionary	Gazetteer	Number	Recurring	Spell:1	Spell:2	Spell:3	(No Features)	Instances	% Correct
	●								●		46	97.8%
	●		●								52,281	97.2%
	●				●						666	96.4%
	●		●	●							14,409	96.2%
		●	●								14,124	93.6%
	●					●					316	93.0%
		●	●	●							3,826	92.5%
		●			●						362	91.7%
	●			●							1,308	90.4%
		●		●							1,266	90.4%
		●				●					232	80.6%
		●					●				1,251	77.9%
		●						●			26	76.9%
	●										941	69.7%
		●							●		13	69.2%
				●							647	64.6%
		●									9,786	62.2%
			●								5,041	43.1%
			●	●							1,832	36.6%
					●						306	17.0%
						●					489	12.3%
								●			9,061	4.6%
										●	23,178	4.0%
							●				37,886	2.8%
									●		5,766	1.0%
Instances	30,886	69,967	91,513	23,288	1,334	1,037	39,137	9,087	5,825	23,178		
Combined	100,853							54,049				
% Correct	96.4%	82.6%	92.1%	89.4%	76.9%	52.2%	5.2%	4.8%	1.9%	4.0%		
Combined	92.2%							4.8%				

Table III
POTENTIAL SUCCESS RATES (PERCENT CORRECT) OF COMBINATIONS OF FEATURES WITH THE CUT-OFF POINT OF THE DECISION LIST (AT THE DOUBLE LINE) LEARNED FROM THE TRAINING SET. (SEE SECTION III-E.)


TAKEN AGAINST STRONG OPPOSITION. ENEMY
ARTILLERY FIRE WAS STRONG

A:TAKE!	[D]	(43.1%)	AGAINST	[D]	(43.1%)
B:T:JcJT	[]	(4.0%)	L.G_II'TST'	[]	(4.0%)
C:T.;I()?2uI	[]	(4.0%)	AGAINST'	[D]	(43.1%)
S:					
O:TAKE!			AGAINST		
T: incorrect			correct		
A:STRONG	[V:3, D]	(97.2%)	OPPOSITION.	[V:2, D]	(93.6%)
B:STRONG	[V:3, D]	(97.2%)	OPPOSiTIOV.	[]	(4.0%)
C:STRONG	[V:3, D]	(97.2%)	OPPOSITION.	[V:2, D]	(93.6%)
S:			OPPOSITION	[S:1]	(2.8%)
O:STRONG			OPPOSITION.		
T: correct			correct		
A:31710	[N]	(17.0%)	ARTILLERY.	[D]	(43.1%)
B:E1ii:Y	[]	(4.0%)	ARTILLERYFIRE	[]	(4.0%)
C:EUEJQY	[]	(4.0%)	ARTILLERY	[D]	(43.1%)
S: ENJOY	[S:1]	(2.8%)	ARTILLERY-FIRE	[S:1]	(2.8%)
O:31710			ARTILLERY.		
T: incorrect			incorrect		

Figure 4. Partial aligned output of the OCR engines from Communiqué No. 204 with assigned features and scores. Hypotheses selected for output are underlined. Numbers in parentheses are assigned weights from Table III. "O:" indicates the word selected for output by the system. "T:" indicates whether the selected hypothesis is correct or not.

Spell:2 features. The plot in Figure 3 shows the relationship between the score and the resulting WER.

If the correctness score of the features of none of the hypotheses in the column exceeded 4.1%, then the hypotheses are excluded from the output. If there is a tie for the highest score, the hypothesis from the first OCR engine in the alignment with that score is selected. All of the selected hypotheses are evaluated against the transcription to calculate the final word error rate for each document. An example of this is shown in Figure 4.

F. Combined Features Results With Training

We applied the method of Section III-E to the OCR output files of the development test set; the results are shown in Table IV under the heading "Combined Features WER". This method has a lower word error rate than the methods using the **Dictionary** or the **Voting** features alone, with a 52.2% relative improvement on the mean word error rate of the three OCR engines and 19.5% relative improvement on the best OCR word error rate. Of particular interest is that the word error rate for five documents matches the lattice word error rate, the lower bound for accuracy based on the evidence from the aligned OCR outputs. Figure 5 compares the system described here using a combination of features and a decision list with our previous work [8] with this corpus, which used the **Dictionary** feature alone. Figure 6 shows that for the majority of documents the WER of the single feature **Voting** system is higher than the WER of this combined features system. Feature combinations help substantially.

	OCR A	OCR B	OCR C	Mean	Lattice	Voting	Dictionary	Combined
	WER	WER	WER	WER	WER	WER	WER	Features WER
Mean	19.88%	30.37%	50.13%	33.46%	12.96%	22.61%	18.70%	16.01%
Minimum	2.27%	1.45%	3.85%	3.39%	0.61%	1.21%	1.21%	0.61%
Maximum	80.68%	111.68%	1,666.42%	602.19%	78.41%	95.45%	86.36%	87.50%

Table IV
COMPARING BASELINE ERROR RATES (COLUMN GROUP ONE) TO RESULTS FROM USING **VOTE** AND **DICTIONARY** FEATURES ALONE (COLUMN GROUP TWO), AND COMBINED FEATURES (COLUMN GROUP THREE).

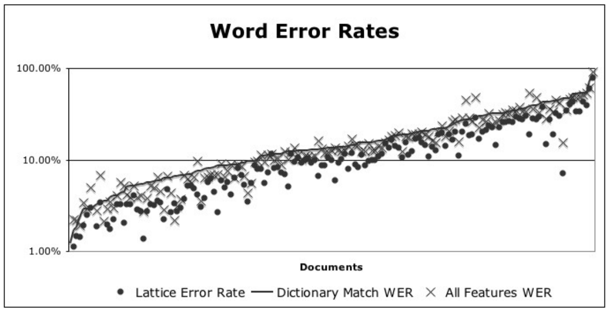


Figure 5. Comparison of the lattice word error rate, Dictionary WER, and error rate using all features.

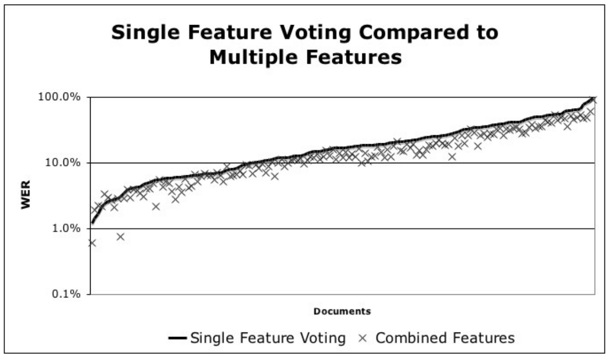


Figure 6. Word error rates from Voting as a single feature versus combined features.

IV. CONCLUSIONS AND FUTURE WORK

This paper presents new insight into the process of using multiple OCR engines and the value in exploiting the variations among engines. We relied on multiple features and trained the scores of the features using success rates on the training set as the basis for a decision list. The resulting method achieved a 52.2% relative improvement over the mean OCR baseline and a 19.5% relative improvement over the mean document WER of the best OCR engine. Further, in a few cases the minimum WER actually matched that of the lattice word error rate. Our goal remains to match and hopefully beat the lattice word error rate.

The underlying problems with some documents will prevent getting anything close to a transcription. Sometimes the true transcription for certain hypotheses is not to be had; however, in regions of the document where good transcriptions can be recovered, those transcriptions can still provide value. Future work should explore the degree to which regions of document image quality can be identified.

In conclusion this paper has explored the use of in-domain training; where in-domain data is not available, techniques for using out-of-domain training data such as those explored in Lund, Walker, and Ringger (2011) [9] can be used.

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] J. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Automatic Speech Recognition and Understanding. Proceedings., IEEE Workshop on*, pages 347–354, 1997.

[4] J. Hull. Incorporating language syntax in visual text recognition with a statistical model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(12):1251–1255, 1996.

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

[6] A. Kae and E. Learned-Miller. Learning on the Fly: Font-free approaches to difficult OCR problems. In *Proceedings of the International Conference on Document Analysis and Recognition*, 2009.

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

[8] 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.

[9] W. B. Lund, D. D. Walker, and E. K. Ringger. Progressive Alignment and Discriminative Error Correction for Multiple OCR Engines. In *Proceedings of the 11th International Conference on Document Analysis and Recognition.*, Beijing, China, Sept. 2011.

[10] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *cs/0010012*, 2000. Computer Speech and Language 14(4), 373-400, October 2000.

[11] Y. Nakano, T. Hananoi, H. Miyao, M. Maruyama, and K. ichi Maruyama. A Document Analysis System Based on Text Line Matching of Multiple OCR Outputs. In *Document Analysis Systems VI*, pages 463–471. Springer-Verlag New York, Inc., 2004.

[12] E. K. Ringger. *Correcting Speech Recognition Errors*. Dissertation, University of Rochester, 2000.

[13] E. K. Ringger and J. F. Allen. A Fertility Channel Model for Post-Correction of Continuous Speech Recognition. In *Fourth International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, PA, Oct. 1996.

[14] R. L. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, 1987.

[15] C. M. Strohmaier, C. Ringlstetter, and K. U. Schulz. Lexical Postcorrection of OCR-Results: The Web as a Dynamic Secondary Dictionary? In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Edinburgh, Scotland, 2003.

[16] G. R. Thoma. Automating the production of bibliographic records for MEDLINE. R&D report, Communications Engineering Branch, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, Sept. 2001.

[17] M. L. Wick, M. G. Ross, and E. G. Learned-Miller. Context-sensitive error correction: Using topic models to improve OCR. In *Proceedings of the International Conference on Document Analysis and Recognition*, 2007.