Evaluation of spell correction on noisy OCR data

Aayushee Gupta, Haimonti Dutta

Abstract

Optical Character Recognition (OCR) of historical text often leads to several kinds of spelling errors. Existing spell correction algorithms do not present a rigorous performance evaluation of the spell correction process. In this paper, we present a novel N-gram based algorithm for evaluation of spell correction which can handle noisy and cleaned text of different lengths. The algorithm relies on appropriately choosing a window of N-words and aligning them in three parallel corpora - noisy OCR, corrected, and cleaned and annotated text (ground truth). Empirical results of spell correction on 300 news articles from the "The Sun" newspaper, Nov- Dec 1894 are presented on edit distance and context sensitive based spell correctors. The Spell Correction Evaluation (SCE) algorithm evaluates their AUC values to be 0.49 and 0.56 respectively. We posit that this novel algorithm for spell correction evaluation has a wide applicability for comparing multiple spell correction algorithms which can play a crucial role in analyzing large volumes of digitized OCR text.

Keywords: OCR, Spell Correction, Spell Correction Evaluation, Historical Newspaper Archives

1. Introduction

OCR of typed, handwritten or printed text is widely used to obtain digitized text which can be edited, searched, stored and displayed efficiently([1, 2]). It is used in various applications such as banking, digital libraries [3] and repositories, number plate recognition, and handwriting recognition [4]. However, the OCR scanning of printed text generates a lot of garbled text which renders them inadequate for any such tasks. /////DOUBT: Should it be mentioned that OCR of specifically historical documents generates garbled text?

Refinement of such noisy OCR text through spell correction can make them useful for text mining tasks ([5], [6]).

//////////DOUBT: Can we say that the USP of our algorithm is that it can be easily used to compare amongst multiple spell correction algorithms and our algorithm will tell which one is suited to a user's application best?

Most spell correction algorithms have focused on improving the correction model and either do not give a detailed performance evaluation of the algorithm post spell correction or the evaluation measures used are not able to completely analyze the performance of such algorithms. A major problem that surfaces when evaluating a spell corrector for OCR text is that the text has to be verified against the original text (ground truth) to estimate its performance. This one-to-one verification may lead to word alignment problems, since the corrected and original text can be of different lengths because of the variety of errors in the OCR text. In this paper, we describe the development of an N-word grams Spell Correction Evaluation (SCE) algorithm that can automatically evaluate a spell correction algorithm by using an N-word window to align three parallel corpora - the noisy OCR, corrected and original/ manually cleaned text. We present the performance evaluation results of applying our SCE algorithm on two commonly used spell correction techniques - Edit distance and Context Sensitive spelliing correction on 300 OCR historical newspaper articles of "The Sun" newspaper published in Nov-Dec 1894.

Organization: This paper is organized as follows: related work is described in Section 2; characteristics of the OCR data and its annotation in Section 3; spelling correction and evaluation algorithms in Section 4; empirical evaluation in Section 5 followed by discussion and future work in Section 7.

2. Related Work

Spelling correction algorithms have been vastly discussed in literature but either their evaluation measures and data analysis are not accurate or they do not take all types of OCR errors into consideration. Kukich[7] comprehensively discusses various spelling correction techniques based on non word, isolated word and real word spelling errors. N-gram analysis, dictionary lookup and probabilistic techniques ([8], [9]) are used for correcting isolated and nonword errors while context-dependent techniques are used mostly for correcting real word errors including the correction of word split and join errors [10]. N-gram techniques work by examining each n-gram in the text string and comparing against a pre-compiled table of n-gram statistics to retrieve the correct word while dictionary look up techniques directly check whether the text string appears in the dictionary using string matching algorithms. Both techniques require a dictionary or a large text corpus and take frequency of n-grams or word occurrence into account in order to find the correct spelling . Probabilistic techniques use transition and word confusion probabilities to estimate likelihood of the correction in order to rank and retrieve correct word spelling. On the other hand, Context-dependent techniques require contextual information and use either extensive NLP techniques or Statistical Language Modeling (SLM) for spelling correction. Bassil and Alwani[11] use Google 1-5 gram word dataset to gain context information in order to determine the correct words sequence in the text for correction. Tong and Evans[12] use Statistical Language Modeling (SLM) approach involving information from letter n-grams, character confusion and word bi-gram probabilities to perform context sensitive spelling correction obtaining a 60 percent error reduction rate. All these spelling correction techniques have developed over time and have been used in combination to achieve improved accuracy [13]. Agarwal et al.[14] use a combination of Google suggestions, LCS and character confusion probabilities for choosing the correct spelling on a small set of historical newspaper data and achieve recall and precision of 51% and 100% respectively.

The edit distance approach, suggested initially by Wagner and Fischer[15], is a dictionary lookup approach commonly used for OCR data correction because of the large number of substitution errors in OCR data [7][16] which can be corrected using this technique. String edit distance approaches with faster correction are discussed in [17],[18] with variants like Levenshtein automata and normalized edit distance. All of the above algorithms are evaluated based on the percentage of spelling errors corrected or reduction in the word error rate and do not consider the word alignment problem arising due to word split and join errors in the OCR text.

Semi-automatic spelling correction systems [19] require user interaction in order to perform complete correction and system evaluation. Rice[20] discusses OCR errors similar to the ones in our dataset. Their algorithm evaluates edit distance spelling correction by estimating word accuracy; the length of LCS between correct and incorrect strings on a page-by-page level is used as the relevant metric. The evaluation strategy works correctly but the definition of accuracy does not give a complete coverage of the spell correction as it does not provide any information on the errors missed by the spelling corrector due to lack of word by word comparison/alignment during the evaluation procedure.

—ADD RELATED WORK REGARDING LINGPIPE BASED CONTEXT SENSITIVE SPELL CORRECTION HERE. LingPipe's model supports exactly this kind of context-sensitive correction. it is able to split tokens as well as combine tokens train up on the data which we are indexing with the search engine. sets the weights for operations like deletions, insertions etc. doing the language modeling at the character level ; no of characters too sample in d ata modeling training is done specific to tokenization Once the model and index are written to disk, we can read them in and correct queries

—ADD RELATED WORK REGARDING RECENT PAPERS READ ON SPELLING CORRECTION EVALUATION. (REYNAERT PAPERS, Compound words paper)

–specifically mention that most of the evaluation mechanisms use OCR and true word pairs for measuring the algorithm performance which is not suitable for a complete evaluation of the spelling correction algorithm where the dataset has word split and join errors.

A lot of research has been done on OCR post correction that use several spell correction models and evaluation measures. But most of them don't consider word split and join errors in the dataset due to which those measures aren't directly applicable on our dataset. OCR and true word pairs are used to measure Word Error Rate for which it is difficult to generate the word pairs when a positional correspondence between the OCR and true word is not possible due to some extra or missing words in the OCR text. For example, [21] consider only unigrams spell correction and don't consider gold standard words involving any spaces

[22] suggest an algorithm for finding OCR and true word pairs but assume that many words are recognized correctly for each line in the OCRed document, and use these words as anchor points for pairing. However, OCR of a historical newspapers generates lots of errors for which this assumption doesn't hold true. (Give an estimation of the number of errors on avg on each line in our ocr documents)

3. Data

3.1. Data Characteristics

The dataset used for empirical evaluation of the algorithm has been obtained from the Chronicling America[1] website. It contains scanned newspaper pages published in New York between1890 to 1920. OCR software is run over high resolution images to create searchable full text of the newspaper articles.

[1] http://chroniclingamerica.loc.gov/

In order to make a newspaper available for searching on the Internet, the following processes used in [5] must take place: (1) the microfilm copy or paper original is scanned; (2) master and Web image files are generated; (3) metadata is assigned for each page to improve the search capability of the newspaper; (4) OCR software is run over high resolution images to create searchable full text and (5) OCR text, images, and metadata are imported into a digital library software program. The scanned newspaper holdings of the NYPL offers a wealth of data and opinion for researchers and historians.The newspaper titles and digitized pages available through the Chronicling America website can be searched using the OpenSearch protocol[2]. Unfortunately, the current search facilities are rudimentary and irrelevant documents are often more highly ranked than relevant ones. The newspapers are scanned on a page-by-page basis and article level segmentation is poor or non-existent; the OCR scanning process is far from perfect and the documents generated from it contains a large amount of garbled text. In a bid to serve its patrons better, the New York Public Library employed human annotators to clean headlines of articles and text, but the process of manually reading all the old newspapers article-by-article and cleaning them soon became very expensive.

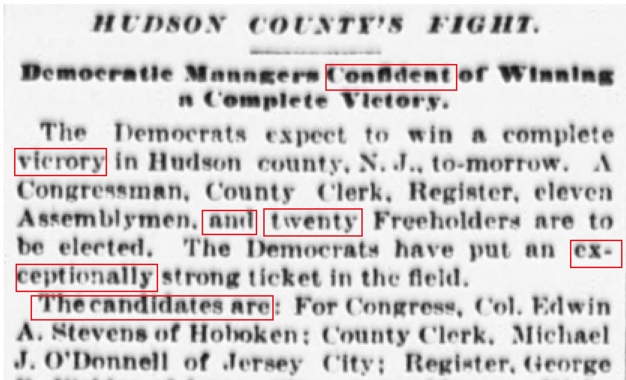
An individual OCR text article has at least one or more of the following types of spelling errors:

- **Real word errors** include words that are spelled correctly in the OCR text but still incorrect when compared to the original newspaper article image. For example: In Figure 1, the word "coil" has been correctly spelled in the OCR text but should have been "and" according to the original newspaper article.

- **Non-real word errors** include words that have been misspelled due to some insertion, deletion, substitution or transposition of characters from a word. For eg. In Figure 1, the word "tnenty" in the OCR text has a substitution error ('n' should have been 'w') which is actually "twenty" according to the original newspaper article.

- **Non-word errors** include words that have been spelled incorrectly and are a combination of alphabets and numerical characters. For example: In Figure 1, the word "4anrliteii" which is a combination of alphabets and number and should have been "confident" as per the original newspaper article.

[2] `http://www.opensearch.org/Home`



HUDSON COUNTY'S FIGHT.

Democratic Managers Confident of Winning a Complete Victory.

The Democrats expect to win a complete victory in Hudson county, N. J., to-morrow. A Congressman, County Clerk, Register, eleven Assemblymen, and twenty Freeholders are to be elected. The Democrats have put an ex-ceptionally strong ticket in the field.

The candidates are: For Congress, Col. Edwin A. Stevens of Hoboken; County Clerk, Michael J. O'Donnell of Jersey City; Register, George

```
I IC1iOV fJT1 11U17
I
Democratic Mnnmcer 4anrliteii or Wlselag
n Complete Ylelorv
The Democrats expect to win n complete
v Icrory In Hudson county N J 1 tomorrow A
Congressman County Clerk RegIster eleven
Ascmblymeii coil tnenty Freeholders nre to
bo elected The Democrats have put nil ex
ceptionally strong ticket In the field
Thernndldntesnra For Congress Col VMnln 1
4ti Steven of Holnken t County Clerk Michael
J 1 ODonnell of Jersey City IlegMer tleorge
```

Figure 1: Scanned Image of a Newspaper article (left) and its OCR raw text (right)

- **New Line errors** include words that are separated by hyphens where part of a word is written on one text line and remaining part in the next line. For example: In Figure 1, the word "ex-ceptionally" where "ex" occurs on one line while "ceptionally" in the next and due to no punctuation in the text, they are treated as separate words in OCR text.

- **Word Split and Join errors** include words that either get split into one of more parts or some words in a sentence get joined to a make a single word. For example: In Figure 1, the word "Thernndldntesnra" in the OCR text is actually a combination of three words "The candidates are" while the words "v Icrory" are actually equivalent to a single word "victory" when compared with the original news article.

3.2. Data Statistics

Our corpus consists of 300 OCR text articles from the "The Sun" newspaper issues of November-December 1894 taken from the Chronicling America website. The text does not have any punctuation and contains a large amount of garbled text containing OCR errors mentioned in Section 3. Spelling correction is performed on this corpus to obtain spell corrected text. We also used manually cleaned OCR text to check and validate the correctness of spelling correction procedure.

Two months of articles of "The Sun" newspaper from November-December 1894 consisting of 14020 news articles with a total of 8,403,844 tokens are used for empirical evaluation.

Due to the word split and join errors and new line errors in the OCR text, the 3 parallel corpora , i.e., OCR text, its spell corrected version and the original correct version need to be aligned so that a word by word correspondence could be made among all versions and the spelling correction algorithm could be evaluated.

3.3. Manual Annotation of Data

For the evaluation of spelling correction on spell corrected data, cleaned/ original newspaper text is required so that a word by word matching can be done to see how well the spelling corrector performs. Due to unavailability of a spell corrected and original correct word pairs list, manual annotation of the OCR text articles needs to be done so that the spelling correction evaluation can be done accordingly. In order to obtain the original cleaned text from the OCR text, we used manual annotators who annotated the OCR articles from their newspaper images. For this purpose, we recruited 5 annotators each of whom were given 50 OCR articles and their corresponding images to annotate. For every OCR article, they had to write the corresponding clean line text called as the original newspaper article version by looking at the OCR article's newspaper image/s. The annotators were also asked not to consider any punctuation in their writing of clean text since the OCR text from Chronicling America is devoid of the same. DOUBT: Does this need to be mentioned here: The OCR text article was given as reference so that the number of linetexts in the correctly written version from newspaper image and the OCR text article remain same. Each annotator was given a different set of articles to work with ranging from November-December issue of "The Sun" newspaper and after the annotation process was complete, one more annotator was asked to verify whether the resulting original article was completely error free or not.

4. Theory

This section first describes the two types of algorithms- Edit Distance and Context Sensitive Spell Correction Algorithms followed by a detailed description of the N-Gram based Spell Correction Evaluation (SCE) Algorithm on which the first two algorithms are tested.

4.1. Edit Distance Spelling Corrector

The Edit Distance algorithm based on Levenshtein distance[23] has been used for spelling correction. It is an isolated word correction technique that uses dictionary based-look up and the distance between strings for matching the text and correcting it. An "edit distance"[3] corresponds to the minimum number of insertions, deletions, and substitutions required to transform one string into another.

4.2. Context Sensitive Spelling Corrector

We use LingPipe's spelling correction algorithm for spelling correction based on context. It is based on a noisy-channel model, which models user mistakes (typos and brainos) and expected user input (based on the data). Mistakes are modeled by weighted edit distances and expected input by a character language model. Spelling corrections are motivated by the indexed dictionary which make the algorithm domain sensitive. The dictionary words are added to the search index and a language model. The language model stores seen phrases and maintains statistics about them. When a word is submitted for correction, the spell corrector looks for possible character edits, namely substitutions, insertions, replacements, transpositions, and deletions, that make the word to be corrected a better fit for the language model. The spell corrector is also able to split and combine tokens during spelling correction due to which Word Split and Join errors can also be corrected by this algorithm.

4.3. Spelling Correction Algorithm Evaluation

For evaluating the performance of spell correction, the raw OCR text and OCR text after application of spelling correction algorithm (corrected

[3]Our edit distance algorithm corrects non-real word spelling errors by making at most 2 operations of insertion, deletion and substitution of letters in the word. The choice of 2 is governed by the trade off between algorithm runtime and quality of spelling correction. The spelling corrector has been designed as suggested by Peter Norvig `http://norvig.com/spell-correct.html`.

text) needs to be compared with the original newspaper text. The OCR text is extremely garbled with Word Split and Join errors due to which word-to-word alignment with the original newspaper text is impossible,i.e., the raw OCR and original newspaper text are of different lengths. A novel algorithm,Spelling Correction Evaluation (SCE) based on N-gram approach is proposed for automatic evaluation of the corrected text. The SCE algorithm can be used for evaluation of any type of spelling correction algorithm - dictionary look up, context sensitive or probabilistic spell correction algorithms. The following metrics are used for estimating the performance: (1) **Accuracy** This requires calculation of the number of OCR errors that got corrected when compared to the original newspaper text. Specifically, $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$ where, TP=Number of True Positives, TN=Number of True Negatives, FP=Number of False Positives, FN=Number of False Negatives. Reynaert and Martin[24] suggest a way to define these terms by distinguishing between correct words and incorrect words in the text through the set of non-target, target and selected words and use Precision and Recall evaluation measures for measuring performance of spelling correction which we adapt for this work. (2) **Time taken for Spelling Correction** The time for correcting the text is also noted for benchmarking correction of large datasets.

N-Word Grams Spelling Correction Evaluation(SCE) Algorithm To make the correspondence between corrected and original OCR text, a window of N-word grams in the original newspaper text is considered which can be seen in a diagrammatic representation in Figure 2. For each token in the spell corrected text, the corresponding token in the original text article along with 2 tokens before and 2 tokens after it are considered for alignment[4]. If the token being considered matches with any word in the original text article word's window and its spelling has been corrected when compared to the corresponding token in raw OCR text, then it is marked as a "True Positive" which is actually rewarding the Spell corrector for making the correct spelling change. A "False Positive" is marked if it does not match any of the words despite its spelling being corrected. Table 1 describes the the process of marking each token in the corrected text as a TP, TN, FP or FN in each text article for calculation of accuracy. The final values of TP, TN, FP and FN are accumulated throughout the dataset to calculate

[4]The choice of N=2 is based on the Word Split and Join errors in the dataset. This value can be set appropriately by considering the maximum difference of lengths in each line of OCR and original text in the dataset.

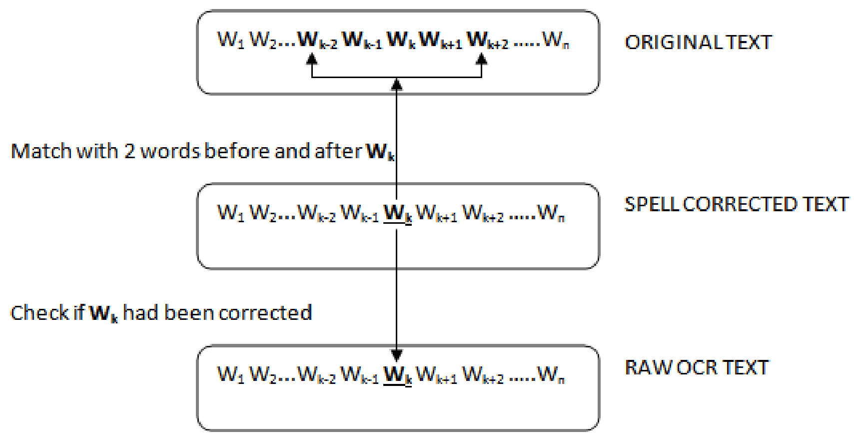


Figure 2: Schematic diagram for alignment of spell corrected article text with original article text for a word W_k

accuracy.

Table 1: Calculation of accuracy in SCE algorithm

Evaluation Metric / Criteria to be checked	TP	TN	FP	FN
Match found in N-word window?	Yes	Yes	No	No
Spell correction was done?	Yes	No	Yes	No

Several scenarios could arise during the word alignment process due to difference in the lengths of text between OCR and original text. All such cases are depicted in Table 2 which describes the window size of tokens to match in the original text (from j=starting index to ending index) for every token i of the corrected text.
A limitation of the SCE algorithm is that it requires all 3 versions of a newspaper article (Original, Corrected and OCR) to have the same number of lines as alignment of line texts is performed. In case of difference in the number of lines of text due to some Word Split and Join errors, the word's window needs to be extended so as to cover previous and next line texts also for alignment.

An Illustrative Example The execution of the SCE algorithm can be demonstrated with the help of the following example: Consider 3 versions of a scanned image of a newspaper article – the original text of the scanned

Table 2: Different cases for word alignment

Token index of OriginalLine / Token index of CorrectedLine(i)	Starting index (j)	Ending index (j)
Length[CorrectedLine] <4 or Length[OriginalLine]<4	0	Length[OriginalLine]
i=0	0	3
i=1	0	4
i=Length[CorrectedLine]-2	i-2	Length[OriginalLine]
i=Length[CorrectedLine]-1	i-2	Length[OriginalLine]
i=Length[CorrectedLine]	i-2	Length[OriginalLine]
i=Length[CorrectedLine+1]	i-2	Length[OriginalLine]
i>=Length[CorrectedLine]+2	Length[OriginalLine]-3	Length[OriginalLine]
Any other value of i	i-2	i+3

image, the raw OCR text and the text after spell correction. Assume, the texts are:

OcrLine= *by tltn rejmrt of th cepert aocountauts who*

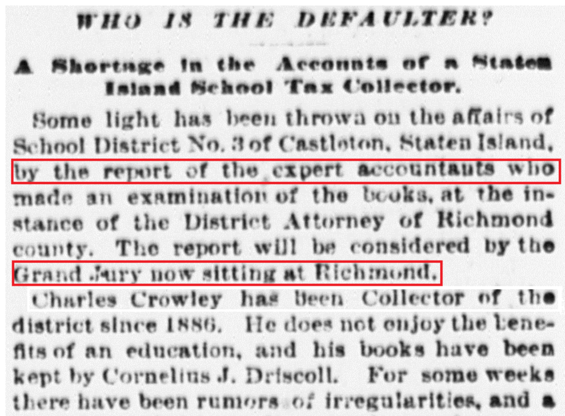
CorrectedLine= *by than report of the expert accountants who*

OriginalLine= *by the report of the expert accountants who*

Here, for each token of CorrectedLine, we find its index and call the MatchWordGrams function accordingly. For the first token 'by' at index i=0 in CorrectedLine, we consider the word window to be "by the report" (index j=0 to 2) in OriginalLine by matching iteratively with each token to see if there is a match and also if there has been a spelling correction by comparing with the corresponding token in OcrLine. Here, no change was made to the spelling of 'by' and it matches with a word in words window, so it is marked as a FN. For the second token 'than' at index i=1, we consider the word window to be "by the report of" (index j=0 to 3) for which there is no match in the window but there has been a spelling correction from 'tltn' to 'than', which implies the correction was wrong and the token is marked as a FP. For the third token 'report' at index i=2, we consider the window as "by the report of the" (index j=0 to 4) in Original Line and find that there is a match in the word window and there has been a spelling correction too from 'rejmrt' to 'report' which makes this token a TP. Similarly, rest of the tokens get marked for each line in the Corrected.txt.

Another example can be considered from Line 10 in Figure 3 and Figure 4 where the number of tokens is different in CorrectedLine and OriginalLine. In such a case, direct alignment between tokens is not possible because of which the words window becomes useful. Here, when the last token

‘Richmond’ of CorrectedLine is considered at index i=3, the corresponding words window becomes “Jury now sitting at Richmond” (index j=1 to 5) for which there is a match in the words windows and corresponding spelling has also been changed from ‘tilchmond’ to ‘Richmond’ which makes it a TP. Had the word window not been considered, the corresponding token at index j=3 in OriginalLine would have been chosen as ‘sitting’ which would have resulted in a FP.


WHO IS THE DEFAULTER?

A Shortage in the Accounts of a Staten Island School Tax Collector.

Some light has been thrown on the affairs of School District No. 3 of Castleton, Staten Island, by the report of the expert accountants who made an examination of the books, at the instance of the District Attorney of Richmond county. The report will be considered by the Grand Jury now sitting at Richmond.

Charles Crowley has been Collector of the district since 1886. He does not enjoy the benefits of an education, and his books have been kept by Cornelius J. Driscoll. For some weeks there have been rumors of irregularities, and a

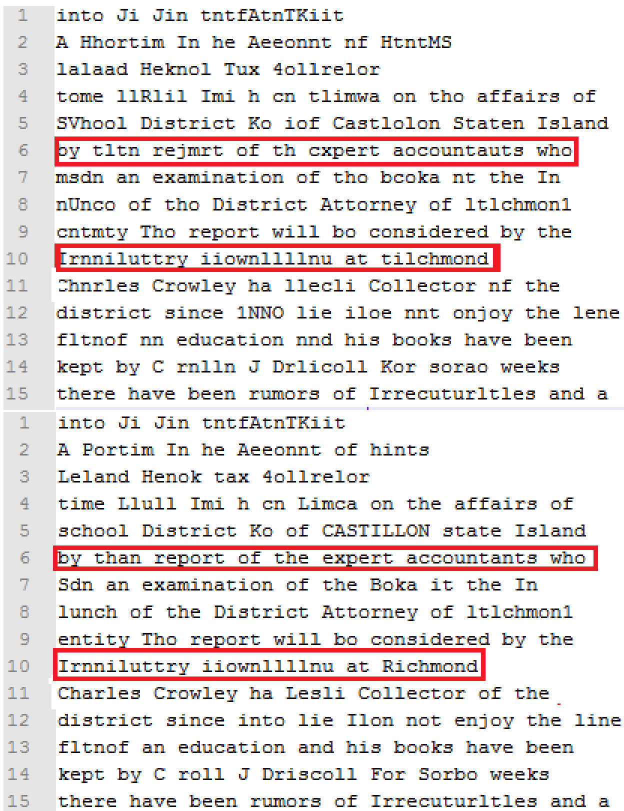
```
WHO IS THE DEFAULTER
A Shortage in the accounts of a States
Island School Tax Collector
some light has been thrown on the affairs of
School District No 3 of Castlaton Staten Island
by the report of the expert accountants who
made an examination of the books at the in
stance of the District Attorney of Richmond
county The report will be considered by the
Grand jury now sitting at Richmond
Charles Crowley has Been Collector of the
district since 1886 He does not enjoy the bene
fits of an education and his books have been
kept by Cornelins J Driscoll For some weeks
there have been rumors of irregularities and a
```

Figure 3: Scanned image of a newspaper article (left) along with its original text (right)

5. Empirical Evaluation and Results

Aim: The aim of our experiments is to answer the following research questions:

RQ1: Which spelling correction algorithm is better?



```
into Ji Jin tntfAtnTKiit
A Hhortim In he Aeeonnt nf HtntMS
lalaad Heknol Tux 4ollrelor
tome 11Rlil Imi h cn tlimwa on tho affairs of
SVhool District Ko iof Castlolon Staten Island
by tltn rejmrt of th cxpert aocountauts who
msdn an examination of tho bcoka nt the In
nUnco of tho District Attorney of ltlchmon1
cntmty Tho report will bo considered by the
Irnniluttry iiownllllnu at tilchmond
Chnrles Crowley ha llecli Collector nf the
district since 1NNO lie iloe nnt onjoy the lene
fltnof nn education nnd his books have been
kept by C rnlln J Drlicoll Kor sorao weeks
there have been rumors of Irrecuturltles and a
```

```
into Ji Jin tntfAtnTKiit
A Portim In he Aeeonnt of hints
Leland Henok tax 4ollrelor
time Llull Imi h cn Limca on the affairs of
school District Ko of CASTILLON state Island
by than report of the expert accountants who
Sdn an examination of the Boka it the In
lunch of the District Attorney of ltlchmon1
entity Tho report will bo considered by the
Irnniluttry iiownllllnu at Richmond
Charles Crowley ha Lesli Collector of the
district since into lie Ilon not enjoy the line
fltnof an education and his books have been
kept by C roll J Driscoll For Sorbo weeks
there have been rumors of Irrecuturltles and a
```

Figure 4: OCR raw text (left) and Spell corrected text (right) of the article

RQ2: How does the variation of the parameters of SCE affects the result?

RQ3: How do parameters of the spelling correction algorithms affect the evaluation results?

Materials: The spelling correction algorithms are used to correct all the 300 OCR raw text articles in the dataset. The dictionary used for look-up is a concatenation of several public domain books from Project Gutenberg and lists of most frequent words from Wiktionary and the British National Corpus[5]. This is augmented with a large people names list which is obtained by running Stanford NER-CRF parser on subsets of the ClueWeb12 dataset made available in the TREC 2013 Crowdsourcing Track[6]. This enhanced

[5]http://norvig.com/big.txt

[6]http://boston.lti.cs.cmu.edu/clueweb12/TRECcrowdsourcing2013/

Table 3: AUC Scores when N=3 for SCE algorithm

ED=2	ED=3	LP
0.491	0.493	0.561

dictionary has been used to give special consideration to correction of person names in the dataset. All our experiments are run on a 4GB RAM machine on Windows platform and the complete implementation of Edit distance and SCE algorithms has been carried out in Java. The context sensitive based spell correction algorithm from LingPipe 4.1.0 [7] version also implemented in Java has been used in our experiments.

//////Mention answer to research questions here

Methods: In order to answer RQ1 for comparing the performance of spell correction algorithms, we use AUC score as a metric as described in previous section. For answering RQ2, we vary the value of the word window N=2 in the N-gram SCE algorithm and compare it with the LCS (Lowest Common Substring) algorithm as well to observe the changes in results of evaluation. For addressing RQ3, we modify the parameters of both spell correction algorithms,i.e., value of edit distance in the edit distance algorithm and other parameters in context sensitive algorithm and study the modified spell correction evaluation results.

Results:

////////DOUBT: Should I mention the time taken to run spell correction algo as a parameter in both the algorithms??

For N-gram SCE algorithm having N=2, the edit distance based spell corrector shows an AUC score of 0.490 when corrected text is compared to OCR text and original article text using our SCE algorithm. We believe that the results are less accurate due to the presence of a large number of non-word, new line, word split and join errors in the OCR data which can not be corrected by the edit distance spelling corrector used for this research. On the other hand, the Lingpipe based context sensitive spelling corrector has a better AUC score of 0.554 demonstrating that in case of OCR text, this algorithm performs better.

We experimented with N=3 for the N-gram based SCE algorithm for which following results were obtained:

[7]http://alias-i.com/lingpipe/

We also compared our N-gram based SCE algorithm with the LCS (Longest Common Subsequence) algorithm[8]. The LCS of corrected and original text gives a list of matching corrected words found in the original text. Following the similar evaluation procedure of calculating accuracy as in the N-word gram approach, if a corrected word finds a match in the LCS and its spelling is found to be corrected, then it is marked as a TP otherwise a TN and if no match is found in LCS and spelling has been corrected, then it is marked a FP or else FN. It was found that there is no statistically significant difference in accuracy when using either of the two algorithms. We posit that LCS is a special case of the N-word gram algorithm when the window size N is set to the total number of words in a linetext.

Our experiments on changing the parameters of spell correction algorithms used in this study reveal no significant changes in accuracy. For Edit distance algorithm, when edit distance is set to 3, the AUC score obtained in 0.492 which is not very high compared to the score obtained for an edit distance of 2. For Context sentitive spell correction algorithm, following parameters were modified:

6. Discussion

The edit distance based spell corrector used in this work corrects non-real word errors by focusing on isolated words in the dataset. We believe a better accuracy of spell correction can be obtained by correcting the new line errors in the articles. This can be done by checking for if the word at last index of a linetext or the word at first index of the next linetext is a word not present in the dictionary and combining the two and checking again in the dictionary for a valid word. The new word, if present in the dictionary can be replaced by the two words from which it is formed thereby removing the New Line error. Similar approach can be applied for word split and join errors but would require each word of an article not present in the dictionary to be analyzed along with some window of words before and after it to make a correction. Since edit distance algorithm is governed by the dictionary choice, using a dictionary with historical terms, places and people names can also help perform spelling correction better and improve its accuracy. The LingPipe based context sensitive spell correction algorithm also employs a corpus dictionary for indexing the terms which suggests that the dictionary choice affects the results of spell correction in this case as well.

[8] https://en.wikipedia.org/wiki/Longest_common_subsequence_problem

7. Conclusion and Future Work

In this paper, we presented a novel approach for automatic performance evaluation of spell correction on noisy OCR text through N-word grams alignment of the OCR, corrected and manually cleaned text. Preliminary results of application of our algorithm on an Edit distance based spell corrector evaluate its AUC score to be 0.49 while it evaluates to 0.56 for Context Sensitive based spell corrector. We believe the N-word gram SCE algorithm can be used to evaluate any kind of spell correction algorithm and provide a detailed performance by considering complete text coverage compared to simpler measures like percentage of corrected words or word error rate used previously. It can be further used to examine multiple spell correction algorithms and analyze the best spell correction algorithm suited for OCR data mining tasks. In future, we plan to add more evaluation metrics to our evaluation algorithm.

8. Acknowledgement

This work has been supported by the National Endowment of Humanities Grant, NEH HD-51153-10.

[1] A. J. Torget, R. Mihalcea, J. Christensen, G. McGhee, Mapping texts: Combining text-mining and geo-visualization to unlock the research potential of historical newspapers.

[2] T. Palfray, D. Hebert, S. Nicolas, P. Tranouez, T. Paquet, Logical segmentation for article extraction in digitized old newspapers, in: Proceedings of the 2012 ACM symposium on Document engineering, ACM, 2012, pp. 129–132.

[3] T. McMurdo, B. MacLennan, The vermont digital newspaper project and the national digital newspaper program, Library Resources & Technical Services 57 (3) (2013) 148–163.

[4] A. Singh, K. Bacchuwar, A. Bhasin, A survey of ocr applications, International Journal of Machine Learning and Computing 2 (3) (2012) 314–318.

[5] H. Dutta, R. J. Passonneau, A. Lee, A. Radeva, B. Xie, D. L. Waltz, B. Taranto, Learning parameters of the k-means algorithm from subjective human annotation., in: FLAIRS Conference, 2011.

[6] T.-I. Yang, A. J. Torget, R. Mihalcea, Topic modeling on historical newspapers, in: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Association for Computational Linguistics, 2011, pp. 96–104.

[7] K. Kukich, Techniques for automatically correcting words in text, ACM Computing Surveys (CSUR) 24 (4) (1992) 377–439.

[8] C. M. Strohmaier, C. Ringlstetter, K. U. Schulz, S. Mihov, Lexical post-correction of ocr-results: The web as a dynamic secondary dictionary?, in: ICDAR, Citeseer, 2003, pp. 1133–1137.

[9] C. Ringlstetter, M. Hadersbeck, K. U. Schulz, S. Mihov, Text correction using domain dependent bigram models from web crawls, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-2007) Workshop on Analytics for Noisy Unstructured Text Data, 2007.

[10] M. A. Elmi, M. Evens, Spelling correction using context, in: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 1998, pp. 360–364.

[11] Y. Bassil, M. Alwani, Ocr context-sensitive error correction based on google web 1t 5-gram data set, arXiv preprint arXiv:1204.0188.

[12] X. Tong, D. A. Evans, A statistical approach to automatic ocr error correction in context, in: Proceedings of the fourth workshop on very large corpora, 1996, pp. 88–100.

[13] E. Brill, R. C. Moore, An improved error model for noisy channel spelling correction, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2000, pp. 286–293.

[14] S. Agarwal, Utilizing big data in identification and correction of ocr errors.

[15] R. A. Wagner, M. J. Fischer, The string-to-string correction problem, Journal of the ACM (JACM) 21 (1) (1974) 168–173.

[16] P. Christen, A comparison of personal name matching: Techniques and practical issues, in: Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on, IEEE, 2006, pp. 290–294.

[17] A. Marzal, E. Vidal, Computation of normalized edit distance and applications, Pattern Analysis and Machine Intelligence, IEEE Transactions on 15 (9) (1993) 926–932.

[18] K. U. Schulz, S. Mihov, Fast string correction with levenshtein automata, International Journal on Document Analysis and Recognition 5 (1) (2002) 67–85.

[19] K. Taghva, E. Stofsky, Ocrspell: an interactive spelling correction system for ocr errors in text, International Journal on Document Analysis and Recognition 3 (3) (2001) 125–137.

[20] S. V. Rice, Measuring the accuracy of page-reading systems, Ph.D. thesis, University of Nevada (1996).

[21] M. W. Reynaert, Character confusion versus focus word-based correction of spelling and ocr variants in corpora, International Journal on Document Analysis and Recognition (IJDAR) 14 (2) (2011) 173–187.

[22] O. Kolak, P. Resnik, Ocr error correction using a noisy channel model, in: Proceedings of the second international conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc., 2002, pp. 257–262.

[23] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, in: Soviet Physics Doklady, Vol. 10, 1966, p. 707.

[24] M. Reynaert, All, and only, the errors: more complete and consistent spelling and ocr-error correction evaluation., in: LREC, 2008.