Evaluation of Spell Correction on Noisy OCR Data

Aayushee Gupta [1]

aayushee1230@iiitd.ac.in

Haimonti Dutta [2]

haimonti@buffalo.edu

[1]Department of Computer Science, Indraprastha Institute of Information Technology, Delhi

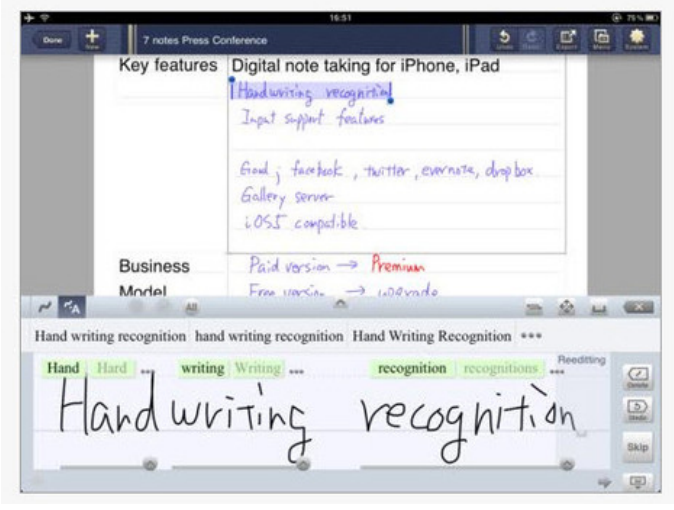
[2]Department of Management Science and Systems, School of Management, University at Buffalo, New York

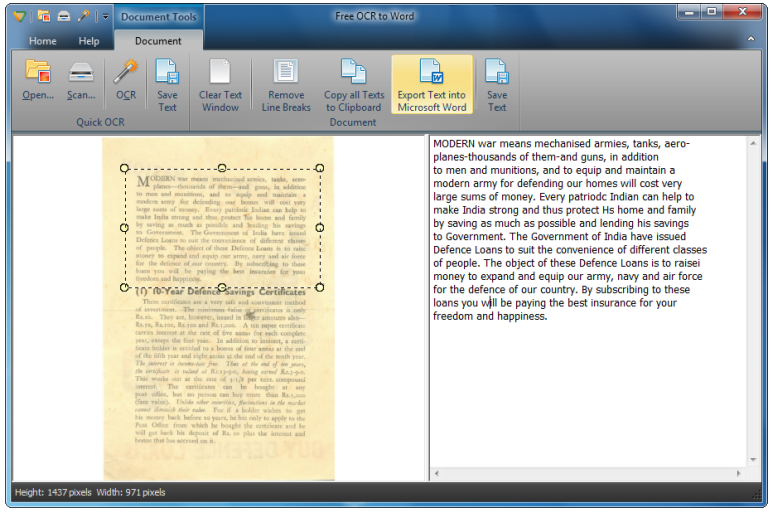
October 31, 2015

Agenda

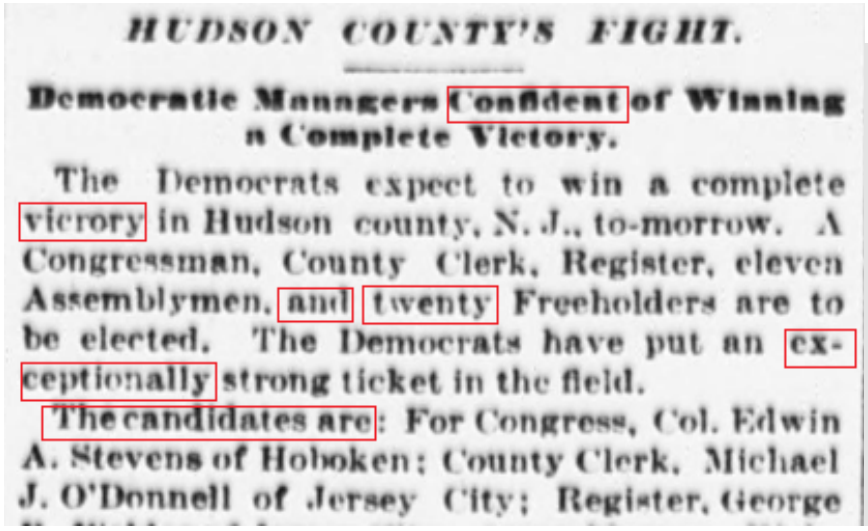
Motivation

- Optical Character Recognition (OCR) is the electronic translation of handwritten, typewritten or printed text into machine translated images
- It has wide applications in the fields of banking, healthcare, digital libraries, handwriting recognition,etc.[7]





An example from digital humanities



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 exceptionally 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

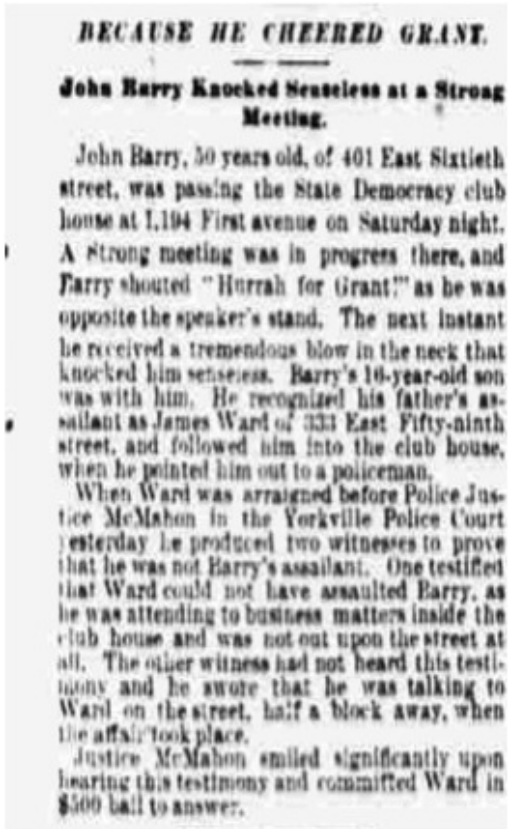
```
1  I ICliOV fJTl llU17
2  I
3  Democratic Mnnmcer 4anrliteii or Wlselag
4  n Complete Ylelorv
5  The Democrats expect to win n complete
6  v Icrory In Hudson county N J 1 tomorrow A
7  Congressman County Clerk RegIster eleven
8  Ascmblymeii coil tnenty Freeholders nre to
9  bo elected The Democrats have put nil ex
10 ceptionally strong ticket In the field
11 Thernndldntesnra For Congress Col VMnln 1
12 4ti Steven of Holnken t County Clerk Michael
13 J 1 ODonnell of Jersey City IlegMer tleorge
```

Figure: Scanned newspaper image and its corresponding noisy OCR text

Motivation

- Spell correction becomes essential as the OCR process generates a lot of noisy text
- Many algorithms exist in literature for automatic spell correction.
- **But how good are they?**
- A major problem that surfaces when evaluating the spell correction process is that the text has to be verified against the original text (ground truth) to estimate its performance.
  - Requires alignment of three parallel corpora - noisy OCR text, text corrected by software, ground truth (often manually verified)

Three Parallel Corpora



BECAUSE HE CHEERED GRANT.

John Barry Knocked Senseless at a Strong Meeting.

John Barry, 50 years old, of 401 East Sixtieth street, was passing the State Democracy club house at 1,194 First avenue on Saturday night. A strong meeting was in progress there, and Barry shouted "Hurrah for Grant!" as he was opposite the speaker's stand. The next instant he received a tremendous blow in the neck that knocked him senseless. Barry's 16-year-old son was with him. He recognized his father's assailant as James Ward of 333 East Fifty-ninth street, and followed him into the club house, when he pointed him out to a policeman.

When Ward was arraigned before Police Justice McMahon in the Yorkville Police Court yesterday he produced two witnesses to prove that he was not Barry's assailant. One testified that Ward could not have assaulted Barry, as he was attending to business matters inside the club house and was not out upon the street at all. The other witness had not heard this testimony and he swore that he was talking to Ward on the street, half a block away, when the affair took place.

Justice McMahon smiled significantly upon hearing this testimony and committed Ward in $500 bail to answer.

JIKCtVSE UK CIIEERKI SIIIXT
John n rry Knoekral Hual le al at Mlroac
MelUc
John Harry M years old of 401 Kast Sixtieth street wa paskInc the htate Dimocrncy club house al 11U4 First avenue on Saturday night A ton rueetlni was In progrcs there and Tar shouted Hurrah for li rant a he I a uplKjjlte the eakerH stand The next Instant he ruclveil H tremendous blow In the neck that knocked him en eies Harrys 11elIln <
MH wih lilm I rrcoicnlifd his father an Millant a James Wall tii Mil East Fiftyninth street and t follourd him Into the cub house when lit pointed him out to n iwllrrnmn 1 When Want wan arraigned before 1ollce Jut tier McMrihim In the Vnrkvllle Police Court Mttteifluy Ie I produied two wltnesea to prove that he was not larn asaallant One teMlfied that I Ward could not hare assaulted Harry as lie I < attendlnt to business matter Inside the i tub untie and was not out tiuoti theMreet at nil Tin other witness liad not heard this teitl titiiiiv and ho sworo that hi was talking to Waril on the ktreet half a block away when IlnMitfaUtofikiilacr
hitlc Aclahnii > mllrd olcnlncantlv ilium lirnrlni Ibis tektlniony and rommitled Ward In 00 bail t answer

JIKCtVSE K CIIEERKI SIIIXT
John n cry Knoekral Had le al at Mlroac
relic
John Harry M years old of 401 last Sixteenth street wa paskInc the state Democracy club house al 114 First avenue on Saturday night A ton rueetlni was In progress there and Tar shouted Hurrah for li ran a he I a uplKjjlte the eager stand The next instant he ruclveil H tremendous blow In the neck that knocked him en eyes Harris 11elIln a
Mr with film I rrcoicnlifd his father an gallant a James Wall tie oil East Fiftyninth street and t followed him Into the cub house when lit pointed him out to n iwllrrnmn 1 When Want wan arraigned before police but tier McMrihim In the Vnrkvllle Police Court Mttteifluy he I produced two wltnesea to prove that he was not learn assailant One teMlfied that I Ward could not hare assaulted Harry as lie I a attendant to business matter Inside the i tub until and was not out tutti theMreet at nail in other witness laid not heard this tell titiiiiv and ho sword that he was talking to War on the street half a block away when IlnMitfaUtofikiilacr
hill Aclahnii a mild olcnlncantlv ilium lirnrlni is tektlniony and committed Ward In 00 bail t answer

Figure: Parallel Corpora: (a) Image of Article (b) OCR (c) After Correction

Types of Errors encountered

- Real words errors: Words that are spelled correctly in the OCR text but still incorrect when compared to the original newspaper article image.
- Non-real word errors: Words that have been misspelled due to some insertion, deletion, substitution or transposition of characters from a word.
- Non-word errors: Words that have been spelled incorrectly and are a combination of alphabets and numerical characters.
- New Line errors: Words that are separated by hyphens where part of a word is written on one text line and remaining part in the next line.
- Word Split and Join errors: Words that either get split into one of more parts or some words in a sentence get joined to a make a single word.

Agenda

Related Work

- Kukich [5] 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 ([1],[3]) are used for correcting isolated and nonword errors while context-dependent techniques([4],[2]) are used mostly for correcting real word errors including the correction of word split and join errors
- All of the above algorithms are evaluated based on the percentage of spelling errors corrected or reduction in the word error rate[6]and do not consider the word alignment problem arising due to word split and join errors in the OCR text

Agenda

Problem Description

Aim: To develop an algorithm that can automatically evaluate a spell correction algorithm so as to align three parallel corpora - the noisy OCR, corrected and original/ manually cleaned text.

Agenda

Components of the Algorithm

- Apply spell correction on the OCR text dataset
- Decide parameters for evaluation of spell correction
- Design an algorithm for spell correction evaluation

Data Gathering

- **Data Source** : Chronicling America - provides scanned OCR newspaper pages of American newspapers published between 1836 and 1922
- **Data Statistics** : 50 news articles of "The Sun" newspaper published between November-December 1894 consisting of tokens
- **Data Characteristics** : News articles consist of one or more OCR errors of the types- Real word, Non-real word, Non-word, Word Split and Join and New line errors, They also do not have any punctuation

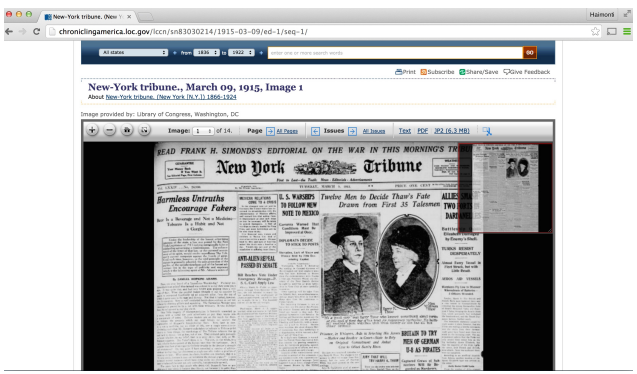


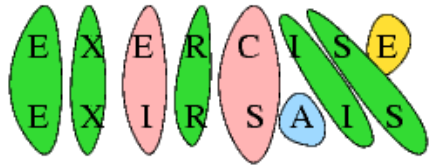
Figure: Chronicling America: A joint effort by the National Endowment of Humanities and Library of Congress to digitize newspapers

Data Preprocessing

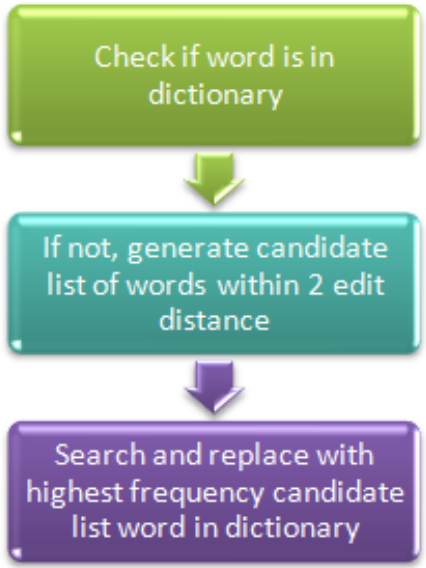
- Required to deal with OCR errors in the news articles
- Edit distance algorithm is used for spelling correction of non-real and non-word OCR errors using precompiled dictionary for look-up
- The dictionary used for look-up is a concatenation of several public domain books from Project Gutenberg and lists of most frequent words from Wiktionary and the British National Corpus augmented with a large people names list extracted from ClueWeb12 dataset

Spelling Correction Algorithm

- "Edit distance" corresponds to the minimum number of insertion, deletion and substitution required to transform one string into another



- String Edit distance algorithm for spelling correction:



- The choice of 2 is governed by the trade off between algorithm runtime and quality of spelling correction.

Spelling Correction Evaluation

- Required to measure the performance of spelling correction

- Evaluation Parameters:

1. Accuracy : measures the percentage of actual errors that get corrected in the OCR text after spelling correction and defined as follows:

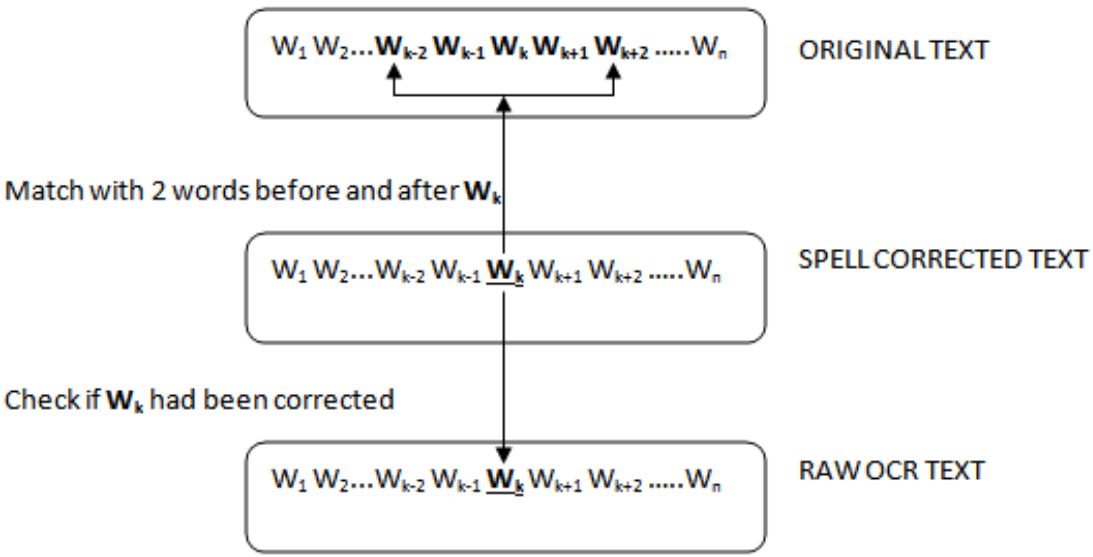
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

where,
TP=Number of True Positives,
TN=Number of True Negatives,
FP=Number of False Positives,
FN=Number of False Negatives.

2. Time taken to run Spelling Correction Algorithm

Spelling Correction Evaluation (SCE) Algorithm

- Word by word correspondence between corrected and original dataset not possible because of Word Split and Join errors in OCR dataset
- SCE algorithm performs word by word automatic evaluation on post spell corrected OCR dataset using an n-word grams approach



Match Found	Spell Corrected	Outcome
Y	Y	TP
Y	N	TN
N	Y	FP
N	N	FN

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

Example

Line text from 3 versions of a news article:

OcrLine= *Irnniluttry iiownllllnu at tilchmond*
CorrectedLine= *Irnniluttry iiownllllnu at Richmond*
OriginalLine= *Grand jury now sitting at Richmond*

Word in Corrected Line	**Corresponding Word Window in Original Line**	**Result**
Irnniluttry	Grand jury now	FN
iiownllllnu	Grand jury now sitting	FN
at	now sitting at Richmond	TN
Richmond	sitting at Richmond	TP

Spelling Correction Evaluation Results

- SCE algorithm tested on 50 spell corrected articles using 3 versions of each article: Original text, Raw OCR text and Spell Corrected text

Accuracy : 73.1%
Time taken : 9 seconds on average per article

- 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.

Agenda

Discussion I

- Spelling Correction accuracy can be improved by correcting other OCR errors like New Line and Word Split and Join errors
- Choice of a dictionary for the edit distance algorithm affects the results of spelling correction
- The choice of window size N=2 in SCE algorithm 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.
- 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 words window needs to be extended so as to cover previous and next line texts also for alignment.

Discussion II

- We compared our N-gram based SCE algorithm with the LCS (Longest Common Subsequence) algorithm 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, 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 complete text in a line.

Agenda

Conclusion and Future Work

- Proposed a novel approach and highlighted challenges for evaluating a spell correction algorithm on noisy OCR dataset 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 accuracy to be 73.1
- SCE algorithm can be used to compare among multiple spell correction algorithms and decide which one suits the dataset better and gives best accuracy
- In future, we plan to use other spelling correction algorithms like context dependent spelling correction to correct the OCR text and measure the accuracy using our SCE algorithm

Agenda

Acknowledgements

This work was initially supported by the National Endowment of Humanities grant no. NEH HD-51153- 10.
The authors would like to thank Barbara Taranto and Ben Vershbow from the NYPL Labs for providing the article level newspaper data and Manoj Pooleery, Deepak Sankargouda and Megha Gupta for setting up the database used in this research.

Thank You.

References

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

[2] Bassil, Y., and Alwani, M.
Ocr context-sensitive error correction based on google web 1t 5-gram data set.
arXiv preprint arXiv:1204.0188 (2012).

[3] Chattopadhyaya, I., Sirchabesan, K., and Seal, K.
A fast generative spell corrector based on edit distance.
In *Advances in Information Retrieval*. Springer, 2013, pp. 404–410.

[4] Elmi, M. A., and Evens, M.
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* (1998), Association for Computational Linguistics, pp. 360–364.

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

[6] Rice, S. V.
Measuring the accuracy of page-reading systems.
PhD thesis, University of Nevada, 1996.

[7] Singh, A., Bacchuwar, K., and Bhasin, A.
A survey of ocr applications.

Table: Different cases for word alignment in SCE algorithm

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

Algorithm 1 MatchWordGrams function of SCE Algorithm for measuring accuracy

```
function MatchWordGrams(OcrLine, CorrectedLine, OriginalLine, jstart, jend, i)
    for (int j=jstart; j<jend; j++) do
        if ((CorrectedLine[i].equals(OriginalLine[j]))&&(!(OcrLine[i].equals(CorrectedLine[i])))) then
            tp = tp + 1  flag0=false  return tp
        end
        else if ((CorrectedLine[i].equals(OriginalLine[j]))&&(OcrLine[i].equals(CorrectedLine[i]))) then
            tn = tn + 1  flag1=false  return tn
        end
    end
    if (!(OcrLine[i].equals(CorrectedLine[i]))&&flag0==true) then
        fp = fp + 1  return fp
    end
    else if ((OcrLine[i].equals(CorrectedLine[i])) && flag1==true) then
        fn = fn + 1  return fn
    end
```