Error Classification in OCR Historic Newspaper Archive using multi-class Support Vector Machine

Megha Gupta and Haimonti Dutta*
Department of Computer Science, IIIT-Delhi
*Affiliated to The Center for Computational Learning Systems, Columbia University, New York
(meghag, haimonti)@iiitd.ac.in

Abstract

Optical Character Recognition (OCR) is a commonly used method of digitizing printed texts so that they can be searched and displayed online, stored compactly and used in text mining applications.
The text generated from OCR devices, however, is often garbled due to variations in quality of the input paper, size and style of the font and column layout. This adversely affects retrieval effectiveness; hence the techniques for cleaning the OCR need to be improvised. Often such techniques involve laborious and time consuming manual processing of data.
This paper shows the need to devise an algorithm that is scalable for a large dataset. The current state of the art algorithm used for performing multi class classification is not yet scalable. The current algorithm takes a long time to converge in a particular parameter setting.

1 Introduction

The *California Digital Newspaper Collection*[1] is an initiative of the Center for Bibliographical Studies and Research (CBSR) which is supported in part by the U.S. Institute of Museum and Library Services. It is also supported by the National Endowment for the Humanities (NEH) to digitise California newspapers for the National Digital Newspaper Program (NDNP). It contains over 400,000 pages of significant historical California newspapers published from 1846-1922.
OCR devices are widely used in electronic conversion of scanned images which are handwritten or printed text into a machine encoded text. The scanning generates It finds most successful applications in the field of machine Learning, Artificial Intelligence and Pattern recognition. It deals with the problem of recognising optically generated characters be it offline or online. The performance directly depends on the quality of input document. The more constrained the input is the better will be the performance of the system. But when it comes to unconstrained handwriting, the performance is far from satisfactory. The main application areas for (Eikvil 1993) like Automatic number plate readers, form readers, signature verification.

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] http://cdnc.ucr.edu/cgi-bin/cdnc

This project deals with printed text in the form of Historical Newspaper Articles in the holdings of (Collection 2009). One such newspaper, The Amador Ledger published in the early 1900s by the Amador Publishing Company appealed to the community's interests by covering issues unique to gold mining. Patrons of the (Collection 2009) continue to be interested in studying about the status of the local mining industry and consequently read the Amador Ledger on a regular basis even to this day and correct (Eikvil 1993) errors as they come across them.
In this paper, we perform error classification using Joachims multi-class Support Vector Machine algorithm called $SVM^{multiclass}$[2]. We chose this algorithm as its the state of the art algorithm till now. But the experiments give altogether a different view on this algorithm. This algorithm do not converge quickly on certain parameters which are shown in table 2

2 Related Work

3 The Data

Log-files

Log-files were generated using a third party software issued by *Veridian*[3], a digital library software. They used this software for logging user text corrections. Using this software, all the corrections made by the patrons were recorded in a xml format log-file.

<TextCorrectedBlock pageOID="AL19000302.1.1" blockID="P1_TB00069" userID="[redacted]">
<TextCorrectedLine lineID="P1_TL00731" >
<OldTextValue>in conjunction willi lln< Calaveras river </OldTextValue>
<NewTextValue>in conjunction with the Calaveras river </NewTextValue>
<OldTextValue>Its Eelation to the Precious </OldTextValue>
<NewTextValue>Its Relation to the Precious </NewTextValue>
<OldTextValue>?; Metal. U </OldTextValue>
<NewTextValue>Metal. </NewTextValue>

[2] http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html
[3] http://www.dlconsulting.com/

</ TextCorrectedLine >
</ TextCorrectedBlock >

The names of the log files follow a convention; the first two letters represent the initials of the newspaper followed by the date in the format yyyymmdd. For example "AL19000105-changes.log" expands to Amador Ledger, 1900-01-05. There were in total 234 log files. To get an idea of the number of corrections made by user per log file, a histogram is shown in figure 1

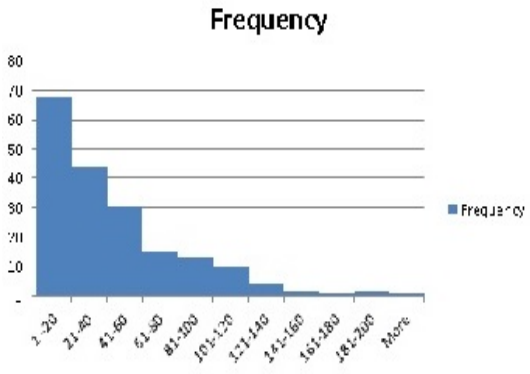

Figure 1: Histogram

The errors rectified by the users were categorised as Spellcheck error , Addition of a new word, Capitalization error, Typo, Punctuation error. The distribution of the these classes in the dataset is shown in figure 2

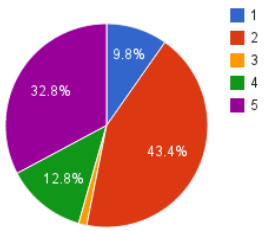

Figure 2: Error Classification

Newspaper

The OCR text generated by running the OCR device on the scanned images of the newspaper was downloaded from the website using Python scripts. The general idea was to decode the names of the log files to retrieve the newspaper name and date. These were further used to formulate URLs needed to download the corresponding OCR text. For instance, "AL19000105-changes.log" was converted to Amador Ledger, 1900-01-05 which was further translated to *http://chroniclingamerica.loc.gov/lccn/sn93052980/1900-01-05/ed-1/seq-1/ocr.txt*.
Here, seq-1 refers to the first page in the sequence of pages in the newspaper. Therefore to extract issues of Amador Ledger, the only thing varied was date. The old and its corresponding new text were stored as key value pairs of a dictionary. Once all the pairs were generated from a log file, the text in the raw corpus matching the keys of the dictionary was replaced with the value corresponding to the matched key.
An analysis was done to figure out the effects of noisy data on text retrieval. The experiments were run on the indexes created on both the datasets using the same query set. The documents retrieved were ranked according to the frequency of the keyword present in the document. Higher the frequency, higher the ranking order. We used Spearman correlation coefficient as the metric to measure the similarity between both our corpora. The average value of Spearman's ranked correlation coefficient calculated by our experiments was 0.625 which denotes that the association between the two corpora is not very strong but the positive value shows that if Raw OCR increases then Corrected OCR will definitely increase.

4 Methodology

We applied state of the art algorithm

5 Empirical Evaluation

Preprocessing & Data Generation

Feature Construction Originally, there were two features in the dataset, that is old text and new text. Further features were manually crafted looking at the types of errors. In our dataset, we have six binary features consisting of sameLength, editDist_0, editDistAbove1, editDistBelow3, editDist_1andcaseChange, punct_diff.

1. "sameLength" is 1 if both the old text and new text have same length
2. "editDist_0" is 1 if both the words are exactly same
3. "editDistAbove1" is 1 if more than one edit operation is required to convert old text to new tex
4. "editDistBelow3" is 1 if less than three edit operations are required to convert old text to new text
5. "editDist_1andcaseChange" is 1 if the two strings have edit distance is exactly 1 and the first character of one string change from upper case to lower case or vice versa.
6. "punct_diff" is 1 if both the old text and new text differ in any of the following punctuation marks !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~

Label Construction The error classes were restricted to 6 classes including Spellcheck Error, Addition of a new word, Capitalization Error, Typo, Punctuation Error and No correction These labels were assigned according to the flow graph as shown in figure

1. Spellcheck error : When the edit distance is between 1 and 3. For example, mounten and mountain.
2. Addition of a new word : When the edit distance is more than 3. For example, at and attend.
3. Capitalization error : When the edit distance of two strings is exactly 1 and first letter of both the strings changes from upper to lower case or vice versa. For example, largest and Largest.
4. Typo : When the edit distance is exactly one and case change is 0. For example, teh and the
5. Punctuation error : When the two strings differ by special characters contained in the set (!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~). For example,
6. No correction : When the old and new text are same. For example, plant and plant

The dataset was parsed to a format used by the Joachim's multi class SVM algorithm which is represented as
<target> <feature>:<value>.....<feature>:<value>
The number of rows in the dataset is 44,022. The class distribution in the dataset is shown in table 1

$Class$	$no.of instances$
1	1970
2	8732
3	261
4	2572
5	6602
6	23885
Total	44022

Table 1: Class Distribution

Experiment Setup

Experiments were performed by randomly partitioning the data into 70% and 30% of training and testing data respectively. For each experiment, the regularisation parameter, C and the type of kernel was varied. The experiments were performed on three machines, two of which were linux servers and the other was a dual core Mac machine with Intel Core i7 processor, 8GB RAM, 2.9 GHz of processor speed. Each experiment was iterated over different dataset for 5 times. The Average Loss and CPU runtime were noted to analyse the experiments.

Results

Discussion

6 Conclusion & Future Work

Acknowledgment

This work is supported by funding from the National Endowment for Humanities. I would also like to express my gratitude to my Advisor, Dr. Haimonti Dutta for her support, patience and encouragement throughout the research.

C	AE_L	AE_P	AE_R	AT_L	AT_P	AT_R
.001	1	2	3	4	5	6
.01	1	2	3	4	5	6
.1	1	2	3	4	5	6
1	1	2	3	4	5	6
10	1	2	3	4	5	6
100	1	2	3	4	5	6

Table 2: Experiment Results

References

America, C. 2009. http://chroniclingamerica.loc.gov/.

Collection, C. D. N. 2009. http://cdnc.ucr.edu/cdnc.

Digital Library Consulting, University of Waikato in Hamilton, N. Z. 1990. http://www.dlconsulting.com/.

Eikvil, L. 1993. Optical character recognition.