A Survey on Various OCR Errors

Atul Kumar
Department of computer Science
Punjabi University Patiala

ABSTRACT

Research has been carried out in correcting words in OCR text and mainly surrounds around (1) non word errors (2) isolated word error correction and context dependent word correction. Various kinds of techniques have been developed. This papers surveys various techniques in correcting these errors and determines which techniques are better.

General Terms

Optical Character Recognition, Natural Language Processing

Keywords

OCR, Errors, NLP. Probability

1. INTRODUCTION

When scanned images of text are converted into machine encoded digital form using OCR, then during conversion there are some errors in the actual text and text converted in machine encoded form during some noisy channels. These errors occur during various stages of OCR like in sentence boundary detection, tokenization and part-of-speech tagging. OCR system fails to recognize a character, an OCR error is produced [3], commonly causing a spelling mistake in the output text. For instance, character “B” can be improperly converted into number “8”, character “S” into number “5”, character “O” into number “0”, and so forth [1]. To remedy this problem, humans can manually review and correct the OCR output text by hand. This task of human interpretation is very time consuming and error prone in nature. There are different techniques that are used to correct these kind of errors. These errors result in non-word errors (2) isolated word error correction and context dependent word correction.

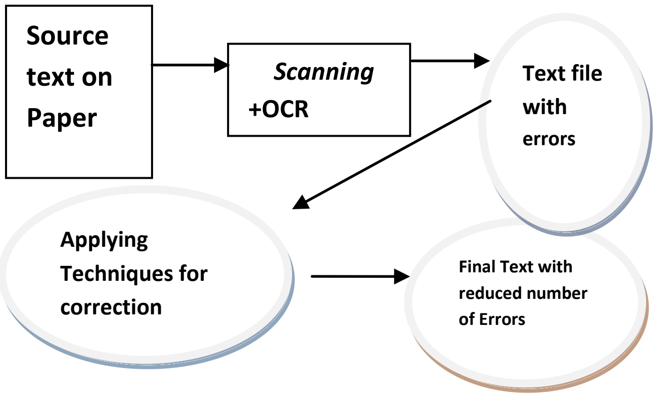
Source text on Paper
Scanning +OCR
Text file with errors
Applying Techniques for correction
Final Text with reduced number of Errors

Fig 1: Complete architecture of OCR software

2. CLASSIFICATION OF OCR ERRORS

Kukich[4]shows three types of errors detection

1. **Non word error detection** detect spelling which results into non words.
2. **Isolated word error correction** correcting spelling which results in non-words
3. **Context dependent error correction and detection** using the context to help detect and correct spelling errors (real word errors)

This classification is obviously not sufficient with respect to OCR errors as it only divides errors into three groups. Furthermore, it is unclear what happens to words which are correct but not contained in any dictionary, e.g. names, outdated terms or (historic) spelling variations. With respect to OCR errors, this classification also lacks of several OCR related aspects. Hence, a better classification is needed here to determine which kind of errors occur. [2]

2.1. Segmentation errors.

Different line, word or character spacing lead to misrecognitions of white-spaces, causing segmentation errors (e.g. thisis instead of this is or depa rtmen t instead of department).

2.2 Hyphenation errors.

Tokens are split up at line breaks if they are too long, which increase the number of segmentation errors (e.g. de-partment).

2.3. Misrecognition of characters.

Dirt and font-variations prevent an accurate recognition of characters which induce wrong recognitions of words (e.g. souiid instead of sound or &-Bi1rd#! instead of Bird).

2.4. Punctuation errors.

Dirt causes misrecognitions of punctuation characters. This means points, commas, etc. occur more often in wrong places with missing or extra white-spaces etc.

2.5. Case sensitivity.

Due to font variations, upper and lower case characters’ can be mixed up (e.g. BrItaIn or BRITAIN).

2.6. Changed word meaning.

Misrecognized characters can lead to new words which are often wrong in context but spelled correctly (e.g. mad instead of sad)

3. TECHNIQUES TO HANDLE THESE ERRORS

There are various techniques for OCR error correction.

1. Dictionary Look up Techniques
2. Key similarity Technique

3. N-grams Techniques
4. Minimum Edit distance techniques
5. Probalististic Techniques.
6. Neural Networks.

3.1.Dictionary Lookup Techniques

Dictionary look up method is a direct method that search for a word in the dictionary for correction. The main problem with this technique is that when the word size exceeds beyond few hundred words. This can be compensated by the use of hash tables. To look up an input string, one simply computes its hash address and retrieves the word stored at that address in the pre constructed hash table [11]. When mismatch occurs in the hash table with the searching word then word is incorrect. Other standard search techniques for searching involves the use of trees like binary search trees where searching is done through calculating keys. If key is less than root node, then search left subtree otherwise right subtree is searched. These are mainly used for non-word error correction.

3.2.Key Similarity Techniques

The main idea behind key search techniques is to map the search string with the strings that has similar keys. Lehal[5] used key search techniques. When a key is computed for a misspelled string it will provide a pointer to all similarly spelled words (candidates) in the lexicon. Similarity key techniques have a speed advantage because it is not necessary to directly compare the misspelled string to every word in the dictionary. The first system developed by this technique was SOUNDEX. It has been used in various applications like airline reservation systems. It involves the match a word in the key consisting of first word followed by digits.

3.3.Rule Based Techniques

Rule-based techniques are algorithms or heuristic programs that attempt to represent knowledge of common spelling error patterns in the form of rules for transforming misspellings into valid words. The candidate generation process consists [9] of applying all applicable rules to a misspelled string and retaining every valid dictionary word those results. Ranking is frequently done by assigning a numerical score to each candidate based on a predefined estimate of the probability of having made the particular error that the invoked rule corrected.

3.4.N-Gram-Based Techniques

N-grams approximate the probability of a word given all the n previous words. There are various types of N-grams like bigrams, trigrams. The probability of word based on the previous word is called MARKOV assumptions (first order). The second order is trigram in which probability depends upon previous two words. The word N-gram approach to spelling correction and detection was proposed [1].The idea is to generate every possible misspelling of each word in a sentence either by typographical modifications and then choose the spelling that gives the highest prior probability using N-gram grammar to compute the probability. There are many statistical approaches to context sensitive spelling corrections. These approaches use Bayesian classifier with trigrams.

In contrast to the information retrieval application, a spelling correction matrix in which words are represented by bigram and unigram vectors [7] are far less sparse is used. Once such a matrix has been decomposed into three factor matrices and reduced, a misspelling is corrected by first summing the vectors for each of the individual- letter n-grams in the misspelled string (as represented in the first matrix) and then multiplying the sum vector by the singular-value matrix of weights (represented in the second matrix). The resultant vector determines the location of the misspelled word in the n-dimensional lexical-feature space. Any standard distance measure (such as a dot product or a cosine distance) can be used then to measure the distances between the vector for the misspelled word and the vectors for each of the correctly spelled words (represented by the third matrix) in order to locate and rank the nearest correctly spelled words.

3.5.Minimum Edit Distance Techniques

It is the minimum number of operations that the one string is converted into other string by means of insertion, deletion and substitutions. For example, minimum edit distance between the and teh is 2 because h is replaced by e and e is replaced by sometimes this technique is also used in postprocessing of OCR to find the candidates whose minimum edit distance is minimum as compare with misspelled word. The minimum edit distance [8] is computed by dynamic programming. Then design a distance matrix for this. Each cell in the distance matrix contain the distance between the first I characters of the target and first j characters of the source. To find the errors, all the strings are searched with minimum edit distance and those becomes the candidates.

3.6.Probabilistic Techniques

The N-grams techniques leads to probabilistic techniques.There are two types of probabilities.Transition probabalities and confusion probabities. Transition probabilities represent probabilities that a given letter (or letter sequence) will be followed by another given letter. Transition probabilities are language dependent. They are sometimes referred to as Markov assumptions. Confusion probabilities are estimates of how often a given letter is mistake as some other letter. Confusion probabilities based on human errors are simply called error probabilities. Let t be the incorrect word and let c be range of candidate corrections.The most likely correction is then

$$\hat{c} = \underset{c\in C}{\operatorname{argmax}} \ \overbrace{P(t|c)}^{\text{likelihood}} \ \overbrace{P(c)}^{\text{prior}} \qquad \text{… (i)}$$

P(t) is constant for all c.

P(c) can be estimated by counting how often the word c occurs in the corpus and then normalizing these counts by the total count ofall the words.

P(c)=C(c)+0.5/N+0.5………………………………(ii)

P(t/c) cannot be computed exactly it can be estimated pretty well, because the most important factors predicting an insertion , deletion , trasposition.One way to estimate these probabilities is the one that Kernighan used.For this confusion matrix[6] , represents number of times one letter was incorrectly used instead of another.A confusion matrix can be computed by hand coding a collection of spelling errors with the correct spelling and then counted the number of times different errors ooccurred.

Using these matrices, estimate P(t/c) as follows

$$p(t/c) \quad = \begin{cases} \dfrac{del\,[cp-1,cp]}{count\ [cp-1cp]} \text{ ,if deletion} \\ \dfrac{ins\,[cp-1,cp]}{count\ [cp-1]} \text{,insertion} \\ \dfrac{sub\,[tp,cp]}{count\ [cp]} \text{ ,if subsitution} \\ \dfrac{trans\ [cp,cp+1]}{count\ [cpcp+1]} \text{ ,if transposition} \end{cases}$$

After this correct word can be estimated by determining multipication of above equation with p(c) by taking the maximum probablity word**.**

3.7.Probabilistic Techniques

The OCR errors can also be corrected using Neural networks[10]. In Neural networks there are various layers based on these layers input to the first layer is applied and the output is produced which generates the number of candidates for misspelled words. There are various algorithms like backpropagation and counter propagation that are used for this. In the backpropagation firstly there is a hidden layer that is used adjust the weights of various layers until correct list is not obtained. There are various activation functions used for this like sigmoid functions etc.

4. CONCLUSION

Based on these techniques the results are shown in table 1

Table 1. Accuracy improved using Different Techniques

Techniques	5000-word dictionary	10000-word dictionary	20000-word dictionary
Dictionary Look up Techniques	64%	67%	54%
Key similarity Techniques	84%	81%	74%
N-grams Techniques	58%	53%	55%
Minimum Edit distance techniques	76%	66%	73%
Probalististic Techniques.	73%	80%	76%
Neural Networks.	88%	82%	84%

It has been concluded that accuracy has been improved using various techniques and most successful techniques are Key similarity and neural Network Techniques.

5. REFERENCES

[1] Bassil, Y., Alwani, M. 2012 . OCR post-processing error correction algorithm using Google's online spelling suggestion. J. Emer. Trends in Computing and Information Sciences. . Res. 3 (Jan. 2012).

[2] Niklas, K. 2010 Unsupervised post-correction of OCR errors. Master's thesis,. Leibniz Universit¨, Hannover.

[3] Lehal, G. S., Singh, C. and Lehal, R. 2001. Shape Based Post Processor for Gurmukhi OCR. In Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01) IEEE Computer Society Press, USA.

[4] Kukich, K. 1992.Techniques for Automatically Correcting Words in Text. ACM Computing Surveys. Res. 24 (Dec. 1992), 377-439.

[5] Sharma, D. V., Lehal G. S. and Mehta S.2009. Shape Encoded Post Processing of Gurmukhi OCR. In proceedings of tenth International Conference on Document Analysis and Recognition.

[6] Yuan, L. X., Chew, L T, Xiaoqing, D., Changsong. 2004.Contextual Post-processing based on the Confusion Matrix in Offline Handwritten Chinese Script Recognition. In proceedings of 17th International Conference on Pattern Recognition ICPR.

[7] Karthika, M., Jawahar, C. V.2010.A Post-Processing Scheme for Malayalam using Statistical Sub-character Language Models. In proceedings of Ninth IAPR International Workshop On Document Analysis Systems, Boston, MA.

[8] Kolak, O. and Resnik, P.2005.OCR Post-Processing for Low Density Languages. In proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.

[9] Bansal, V. and Sinha, K. M. R.1999.Partitioning and searching dictionary for correction of optically read Devnagri character strings. In Proceedings International Conference on Document Analysis and Recognition.

[10] Chaudhuri, B. B., Pal, U. 1998.A Complete Printed Bangla OCR systems. Pattern Recognition.1998. Res. 24 (Mar. 1998), 531-549

[11] Kernighan, M. D., Church, W. K. and Gale, A. W.1990.A Spelling Correction Program Based on a Noisy Channel Model. In Proceedings of the 13th conference on Computational linguistics.