Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 03, March -2018

Survey of Post-Processing Methods For Cursive Handwriting Recognition System

Ann Jisna James(*Author1*), Prof.Reena Kharat (*Author2*)

[1]*Department of Computer Engineering, P.C.C.O.E,Pune*
[2]*Department of Computer Engineering, P.C.C.O.E,Pune*

Abstract— *A vast amount of data to this date is present in physical format stored in handwritten or printed formats.Such data are susceptible to be lost over a long period of time due to various factors such as the storage material,climate etc.Also the form of data is not widely available and information in such case is restricted to the geographical arena in which this material might be available.Hence the need to digital this information for its proper storage as well as wider reach to community . OCR has been profoundly used for this purpose,which converts the image of handwritten or typed text into machine encoded.Therefore a need arises to recognize the right words and discard the incorrect after this process to obtain meaningful and authentic data.Hence post processing methods for error correction needs to be placed in the system.The following paper makes an attempt to survey the present error correction methodologies.*

Keywords— OCR(Optical Character recognition),Post Processing*;*

I. INTRODUCTION

With the on-going digitisation efforts, OCR software constitutes the key to providing unprecedented electronic access to vast volumes of textual data, which were previously only available in printed form, and thus potentially difficult to obtain. Also OCR has found a platform to recognize prescription given by doctors as a recent set of studies have shown that thousands of people die annually because of poorly interpreted medical prescriptions.hence now medical sector has also induced in this system for better medical care.. The increasing market share achieved by so-called ‘pen computers’ indicates that on-line character recognition technology has also reached a high level of maturity. However, many problems within the OCR field remain areas of active research, one of the most important and challenging being that of off-line cursive script recognition. Some of the many applications of off-line recognition technology are automatic mail sorting, bank cheque and forms processing, and the transfer of databases that currently exist only in handwritten form to an electronic format.

However, the efficiency of such conversions can be hampered by the substandard quality of the text recognised by OCR software. Specific features of documents and handwritten data, which can include poor and/or variable print quality/handwriting, unusual fonts or archaic vocabulary, can present major challenges to OCR systems designed to analyse modern documents. Poor OCR output can affect all types of functionalities over the data. For example, if many words are misrecognized, then keyword-based searches may to fail to retrieve potentially relevant documents. More sophisticated search methods needs to be in cooperated.

II. LITERATURE SURVEY

A. Automatic Error Detection and Correction of Text- The State of the Art

The following paper by Zhang yang and Zhao Xiaobing classifies post processing errors into two sets isolated-word error and real-word error.

A. Isolated-word error

Larger proportion of English text errors are isolated-word errors that are not related to the context.Techniques for automatically detecting isolated-word errors are mainly based on N-gram model and dictionary.

The method based on N-gram model, which detects errors mainly through finding the abnormal strings, requires a prior N-gram table edited according to a large-scale text corpus or a dictionary. The text inputted is scanned in every n-string and the n-sting with low frequency or nonexistent is judged as misspelled. The method based on dictionary that has a highly checking accuracy is widely used to detect isolated-word errors with the improvement of computer storage and calculation capacity. This method detects the input strings, the string which does not exist in a dictionary is judged as misspelled.

isolated-word strings and candidate word strings to correct text errors. The representatives are the minimum edit distance approach, similarity key approach, N-gram model-based approach, rule-based method and so forth.

B. Real-word error

Real-word error includes not only spelling errors (such as lord and lore), also common grammatical errors (such as has and have), word boundary confusion errors (such as everyday and every day) and so on. Real-word errors take up a small proportion in English text errors. However, because to detect and correct this type of error rely on the context, the process is more difficult and related to the natural language understanding research. English text really wrong word proofreading methods include rule-based and statistical two categories. The approaches for detecting and correcting real-word error can be based on rules and statistics.

The automatic Chinese text error detection and correction approaches can be classified into three types:

(i) Rule-based approach

This approach uses the grammar rules, phrases rules, etc. of linguistic to check text errors. error combined the pattern matching and sentence component analysis methods.Due to that the limited rules can hardly cover the large number of linguistic phenomena, the results of the rule-based approach of Chinese text error check and correction are unsatisfying.

(ii) The approach using the contextual features

A confusion set for proofreading is created for text. The system finds a candidate word for each word in the text to be proofread. All the candidate words constitute a candidate matrix. Then T-rules is implemented for binding and pruning candidate words to form the language structure elements, from which, it applied Markov model to find the best path. This method takes advantage of the language own structure and statistical features.

(iii) Statistics-based approach

Most text errors lead to disperse strings after segmentation. Put the words that are similar in pronunciation, shape, meaning or code together. Each word in the text be proofread is replace by the ones in the similar character set to form several candidate strings. Then all the candidate strings are scored based on bi-gram model. The one which get the highest score is considered as the correct strings. This method can only detect and correct the wrongly written characters, but difficult to detect the missing words or nslocation, etc.

(iv) The combination approach of rules and statistics

A word matching and syntax analysis combining method to automatic error detection and correction of text. It based on reverse maximum matching algorithm and statistics to find out the disperse strings, which are analyzed according to word matching and syntax analysis. The candidate error strings are corrected by the way of human-computer interaction.

The proposed method takes a set of error detecting rules based on the law of single characters which appear after segmentation in a correct text and which appear due to input error. Then built bi-gram and tri-gram models for the single characters. The final step is combining the rules and models to detect text errors.

B. Effective Spell Checking Methods Using Clustering Algorithms

The authors Renato Cordeiro de Amorim and Marcos Zampieri has proposed a methodology to reduce the number of times distances have to be calculated when finding target words for misspellings. The method is unsupervised and combines the application of anomalous pattern initialization and partition around medoids (PAM).

A.Anomalous Pattern Initialization and PAM

The partition around medoids (PAM) algorithm divides a dataset Y into K clusters S = {S1, S2, ..., SK}. Each cluster Sk is represented by a medoid mk. The latter is the entity yi ∈ Sk with the smallest distance to all other entities assigned to the same cluster. PAM creates compact clusters by iteratively minimising the criterion below.

$$W(S,M) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v \in V} (y_{iv} - m_{kv})^2,$$

where V represents the features of the dataset, and M the returned set of medoids {m1, m2, ..., mK}. This criterion represents the sum of distances between each medoid mk and each entity yi ∈ Sk. The minimisation of above equation follows the algorithm below

1. Select K medoids at random from Y , M = {m1, m2, ..., mK}, S ← ∅.
2. Update S by assigning each entity yi ∈ Y to the cluster Sk represented by the closest medoid to yi . If this update does not generate any changes in S, stop, output S and M.
3. Update each medoid mk to the entity yi ∈ Sk that has the smallest sum of distances to all other entities in the same cluster. Go back to Step 2.

PAM is a very popular clustering algorithm and it has been used in various scenarios..

C. Customised OCR Correction for Historical Medical Text

The paper proposes a Modifying Hunspell.

The scheme follows an approach which changes the basic Hunspell ,which is a spellchecker it has an attached dictionary from which it matches words and retrieves the appropriate result..

The following features has been included:

Firstly, the use of archive based word frequencies is extracted to determine not only how a word should be corrected, but also whether it should be corrected at all. It is assumed that, within the archive, a given word will be recognised correctly by the OCR software more frequently than it is misrecognized. Based upon this,it is assumed that words that are flagged as

erroneous by Hunspell, but whose decade-based archive frequency is greater than the frequencies of any of the suggested corrections, constitute valid words, which is not present in Hunspell's dictionary leaves such words uncorrected.

Secondly, Hunspell's default dictionary has been augmented with the open-source OpenMedSpel dictionary , containing around 50,000 specialised medical terms. This would help reduce the cases in which specialised vocabulary is flagged as erroneous by Hunspell. However, given that OpenMedSpel contains modern medical terms, and since medical vocabulary over the complete archive is subject to change, it is considered useful to combine the use of the augmented dictionary with the "selective" spelling correction introduced above. This helps to prevent erroneous correction of valid words that do not appear even in the augmented Hunspell dictionary.

D. Deriving Symbol Dependent Edit Weights for Text Correction - the Use of Error Dictionaries

The paper has introduced a new method for obtaining symbol dependent edit weights for text correction tasks that does not need any ground truth training data. The method is based on annotated error dictionaries, which is a design of an error dictionary for estimating frequencies of edit operations in an erroneous text is a delicate matter. The methodology uses a simple brute-force construction, based on a set of typical OCR errors. After a series of tests by the following construction was drawn:

(a) To the 100,000 most frequent words of English dictionary (base dictionary) we applied all substitutions, a selected list of "typical" mergers and splits, and deletion of a letter i, l, t, or f. (b) To the 25,000 most frequent words all (other) merges were applied

(c) To the 5,000 most frequent words all (other) splits were applied. For each token of the base dictionary the set of error transformations was sequentially applied at each position. Hence, entries of the error dictionary E contain just one error. For each error W obtained , the retranslation was only stored that leads to the most frequent correct word W.

At this point, when garbling a word with one of the non-typical merges and splits (cases b, c), a reduction in its frequency by a penalty factor of 1/100 is implied. Deletion of all erroneous tokens with a length $\leq$ 3 since acronyms and special names of length $\leq$ 3 can easily be misinterpreted as errors. Also exclusion of all errors that correspond to some word in a large collection of standard dictionaries.

Since each correct word that is left in the error dictionary may lead to a misclassification, it is important to use a collection of dictionaries with very high coverage.

The main advantage of this method is its simplicity and efficiency. The English error dictionary is stored as a finite state automaton .The above experimental results in the area of OCR correction shows that the method based on error dictionaries leads to a significant improvement of correction accuracy. Only a small additional gain is obtained when using ground truth training data. The use of error dictionaries is promising in situations where

(1) a class of error patterns can be specified in advance which captures all frequenterror types

(2) errors found in the text not only affect exotic words with very low frequency.

Assumption:

(1) Most typing errors, for example, can be traced back to letter transpositions and a small set of keyboard specific substitutions and insertions. In these operations, only letters neighbored to the correct letter on the keyboard have to be taken into account.

If condition (2) does not hold, estimating edit weights with error dictionaries becomes more problematic. For the generation of suitable error dictionaries find a base vocabulary of correct words that captures a relevant part of the corrected version of the errors found in the text.The results in the paper results indicate that specialized dictionaries crawled in the web may meet this requirement.

E. OCR Error Correction Using Character Correction

and Feature-Based Word Classification

The author makes an attempt to classify and identify errors through:

1) Feature Extraction:

The features were extracted at the word level:

Confusion weight – The weight attribute of the corruption-correction pair in the confusion matrix, which is the number of occurrences of this pair calculated by the noisy channel on the training set. This feature reflects the OCR error model in accordance with the OCR engine performances over the document images training set, generally affected by font characteristics and scan quality.

Unigram frequency – The unigram document frequency, providing a thematic domain and language feature independent of adjacent words or document context. Backward/Forward bigram frequency – The maximal document frequency of the bigram formed by a correction candidate and any candidate at the preceding/following position. This feature is valuable as it contains an intersection between language model and domain context, but is non-existent for many of the bigrams and is redundant if one of the unigrams does not exist. Although the bigrams should have been calculated in comparison to all the correction candidates, it was taken only on the OCR output due to calculation complexity and the relative rarity of sequential word errors. Furthermore, we set a cutoff frequency to overcome performances issues in the extraction stage

F. Fast string correction with Levenshtein automata

The Levenshtein distance between two words is the minimal number of insertions, deletions or substitutions that are needed to transform one word into the other. Levenshtein automata of degree n for a word W are defined as finite state automata that recognize the set of all words V where the Levenshtein distance between V and W does not exceed n. We show how to compute, for any fixed bound n and any input word W, a deterministic Levenshtein automaton of degree n for W in time linear to the length of W. Given an electronic dictionary that is implemented in the form of a trie or a finite state automaton, the Levenshtein automaton for W can be used to control search in the lexicon in such a way that exactly the lexical words V are generated where the Levenshtein distance between V and W does not exceed the given bound. This leads to a very fast method for correcting corrupted input words of unrestricted text using large electronic dictionaries. We then introduce a second method that avoids the explicit computation of Levenshtein automata and leads to even improved efficiency. Evaluation results are given that also addressvariantsofbothmethodsthatarebasedonmodified Levenshtein distances where further primitive edit operations (transpositions, merges and splits) are used.

III. RESULT

The above survey leads to the following conclusion:

The dictionary method is only useful if a huge repository of words is present; this may also lead to ambiguity

The chances of making a false suggestion are high in case frequent word count is not maintained.

The cluster format may lead to a faster word recognition.

Using language structure through contextual and grammatical format may be helpful.

Dictionary related to the particular arena should be included so as to obtain meaningful correction. eg:medical domain error correction require repository containing medical terminology Error dictionaries are helpful in faster implementation and error checking but they also may lead to incorrect suggestion or correction or can miss correction in case of absence of the particular error from the dictionary Identifying the type of error may be helpful.

CONCLUSION

Based on these techniques the results are shown in table 1

TABLE I. COMPARISON

S. no	***Techniques***	***Working***	***OCR error***
1	Dictionary Look up Techniques	at search for a word in the dictionary for correction,use of trees like binary search trees where searching is done through calculating keys	isolated-word
2	Key similarity Technique	map the search string with the strings that has similar keys	isolated-word
3	N-grams Techniques	N-grams approximate the probability of a word given all the n previous words	isolated-word
4	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.	isolated-word
5	Probalististic Techniques	probabilities are estimates of how often a given letter is mistake as some other letter	isolated-word error ,Real-word error

a.

REFERENCES

[1] Zhang Yang, Zhao Xiaobing "Automatic Error Detection and Correction of Text The State of the Art ", 2013 6th International Conference on Intelligent Networks and Intelligent Systems. .

[2] Renato Cordeiro de Amorim,Marcos Zampieri "Effective Spell Checking Methods Using Clustering Algorithms",Proceedings of Recent Advances in Natural Language Processing September 2013.

[3] Paul Thompson, John McNaught and Sophia Ananiadou "Customised OCR Correction for Historical Medical Text ", 2016 IEEE.

[4] Ch. Ringlstetter,U. Reffle,A. Gotscharek "Deriving Symbol Dependent Edit Weights for Text Correction -The Use of Error Dictionaries ", Document Analysis and Recognition, 2007.

[5] Youssef Bassil, Mohammad Alwani, "OCR Post-Processing Error Correction Algorithm Using Google's Online Spelling Suggestion ",Journal of Emerging Trends in Computing and Information Sciences.

[6] S Procter, J.illingworth and F.Mokhtarian , "Cursive handwriting recognition using hidden Markov models and a lexicon-driven level building algorithm ",IEE Proc.-Vis. Image Signal Process., August 2000.

