User-defined expected error rate in OCR postprocessing by means of automatic threshold estimation

J.Ramon Navarro-Cerdan Joaquim Arlandis Juan-Carlos Perez-Cortes Rafael Llobet
Instituto Tecnologico de Informatica Universidad Politecnica de Valencia Camino de Vera s/n, 46071 Valencia, Spain {jonacer, arlandis, jcperez, rllobet}@iti.upv.es *

Abstract

In this work, a method for the automatic estimation of a threshold that allows the user of an OCR system to define an expected error rate is presented. When the OCR output is post-processed using a language model, a probability, a reliability index (or a “transformation cost”) is usually obtained, reflecting the likelihood (or its inverse) that the string of OCR hypotheses belongs to the model. Using a threshold on this index (or cost) to reject the less reliable hypotheses, a variable level of expected accuracy can be imposed on the output. It is much more convenient for the user the ability to “fix” at an acceptable level the expected error rate instead of having to deal with an arbitrary threshold. Of course, the result will always be high reject rates for difficult tasks and lower reject rates for easier tasks.

1 Introduction

In an Optical Character Recognition (OCR) task on handwritten documents, the quality of the results depends on the knowledge of the task that can be embedded in the system. This knowledge is often encoded in an Error Model and a Language Model that are applied as a post-processing stage. The first one can be more or less fixed since it depends mainly on the OCR classifier and in the typical behaviour of the writers. The second one reflects the constraints that can be imposed on the contents of the text. They can be constraints from the syntax of a natural language, a subset of a natural language, a closed list of words or expressions, a code following some pattern, etc.

In the typical case of forms with pre-defined fields that are filled-in by hand each field holds different information and a model can be often defined for each one. For example, usual fields in commercial forms are “Name”, “Age”, “Date”, “Country”, “Street”, “Phone Number”, etc. The language models associated with each of these fields are often very different in alphabet, size, complexity, perplexity,... and new language models appear routinely in a normal form-processing industrial activity.

There are several works of using language modeling techniques for error correcting applied to OCR and text recognition tasks, either on constrained or unconstrained environments. Some examples can be found in [8], [18], [13], [10]. Most of them provide an estimation of the effort needed to make the output from the OCR classifier comply with the language constraints. This estimation is called transformation cost.

Applying a threshold to this cost allows the system to reject those strings that are less likely to be correct. Usually, the rejected strings are submitted to a manual data-entry process and therefore the threshold selection has a significant impact in the practical performance and economic impact of the system. The maximum acceptable error rate in the accepted strings (which could be regarded as false positives) depends on the particular task at hand, and the number of rejections must be minimized due to the high cost of each manual data-entry operation.

Although the number of rejections is difficult to control, being highly dependent on factors such as the quality of the handwriting style, scanning process, image enhancement, field registration, etc., our hypothesis is...
that the error rate can be predicted for a given language model observing the distribution of the transformation costs of a representative training sample.

But there are a number of issues that make this prediction difficult. Given a language model, different test samples, written by diverse kinds of users or acquired under different conditions, may show very distinct distributions of transformation costs, and therefore require very different reject thresholds to give rise to the same error ratio, as can be shown in Section 5.

In this paper, a technique to estimate the rejection threshold in a sample of transformation costs from a user-defined expected error rate is proposed and experiments are presented comparing the accuracy of the estimation in different conditions.

The rest of the paper is organized as follows: section 2 contains an overview of the related work. Sections 3 and 4 describe how the error probabilities of a set of transformation costs can be used as a model to compute the rejection threshold for a given error rate. In section 5, experiments and results on error rate estimation for different languages and test sets are reported, and, finally, the conclusions are presented in section 6.

2 Related work

Given the practical importance of a good control of the rejection strategy by the user of an OCR system, many works have addressed this issue. A recent study is, for example, the one presented by Landgrebe [11] proposing a modified version of ROC curves, where a factor to tune the number of false positives expected is introduced in order to tackle with imprecise environments. Other works that propose different rejection strategies are [3], [5], [14]. The use of predictable confidence measures and rejection criteria has also special relevance in the field of natural language processing and in speech recognition tasks.

The rejection threshold estimation has also been extensively studied in biometrics, for instance in the task of automatic speaker verification. In [1], several methods for estimating speaker-independent and speaker-dependent decision thresholds were compared using only relevant parameters estimated from training data.

In other disciplines, the same problem arises often. In [6], regression is used for threshold estimation, in sensor systems, where large amounts of data are usually available, target detection is seriously affected by false positives and therefore, the control of the operational point is important. Ozturk et al. [12] used the generalized Pareto distribution to approximate the tails of the distributions of radar measures, and propose the ordered sample least squares method for estimating the threshold. Recently, Broadwater and Chellappa [4] proposed an algorithm using extreme value theory through the use of the generalized Pareto distribution, too, and a Kolmogorov-Smirnov statistical test, and incorporate a way to adaptively maintain low false positive rates minimizing the differences between the model assumptions and the real data.

In handwritten numeral recognition, He et al. [7] used Linear Discriminant Analysis to determine the rejection threshold by taking into account the confidence values of the classifier outputs and the relations between them. In text correction, Kae and Huang [9] used a technique for identifying a set of correct words by bounding the probability that any given word from an OCR output is incorrect using an approximate worst case analysis.

Some works presented by Schlapbach et al. [15], [16] have recently addressed relevant topics related to error estimation of handwriting recognition systems. In those works, they pose a two class classification problem to pre-classify text as either readable or unreadable, based on some image-level feature vectors. This is carried out prior to the use of the text recognizer, and allows the rejection of text in early stages of the process, as well as the estimation of the expected recognition rate of the system.

In the context of real tasks, setting an automatic rejection threshold from an user-defined expected error rate would alleviate the problem of the management of confidence measures. In this sense, a closer goal to which is presented in this work has been tackled by Serrano et al. in the context of error supervision in interactive-predictive handwriting recognition [17]. Such a goal was to assist the user in locating possible transcription errors: the user decides on a maximum tolerance threshold for the recognition error (after supervision), and the system adjusts the required supervision effort on the basis of an estimate for this error.

For OCR systems, Arlandis et al. have recently presented a work involving the estimation of the expected error rate distribution of an unknown language model from a training set composed of known language models using regression techniques [2].

3 Modeling the error rates

Given a language model and a set of transformation costs obtained using a post-processing algorithm with a representative sample of OCR hypothesis, a smoothed histogram \( H_n(c) \) of error rates as a function of the cost \( c \) (the relative number of strings with a cost in the vicinity of \( c \) that are “erroneously corrected” into an incorrect string), can be computed using the expression,
The number of costs around a size can also be defined dynamically to enclose a given number of strings having a cost also in that interval. The window size can also be defined dynamically to enclose a given number of costs around a given size instead of a fixed cost interval.

In figure 1, a histogram \( H_e \) obtained using a language model of Spanish first names (see Section 5 for a detailed description) and the parsing algorithm of [13] to post-process the OCR output of a sample of handwritten Spanish names is shown.

For a given language model, we can use this histogram as a source of information to decide the appropriate cost threshold to use when we want to fix the expected error rate, as discussed in the next section.

\section{Automatic rejection threshold estimation}

Given a specified error rate \( \epsilon \), a rejection threshold \( \mathcal{T}_c \) can be computed for a test sample from a cumulative averaged version of \( H_e \). Given a test sample of transformation costs \( T \), corresponding to a set of OCR hypotheses (strings) sorted according to increasing values of \( c \),

\[ T = \{ c_1 \ldots c_i \ldots c_n \}, c_1 \leq c_i \leq c_{i+1} \leq c_n \, , \]

an approximation of the error rate for each cost \( c_i \) in \( T \) can be computed as a cumulative averaged version of the histogram \( H_e \) as follows,

\[ E(i) = \sum_{c=c_1}^{c_i} \frac{H_e(c,w)}{i} \, , \quad (2) \]

and the rejection threshold \( \mathcal{T}_c \) that corresponds to a given expected error rate \( \epsilon \) is obtained as,

\[ \mathcal{T}_c(\epsilon) = \max_{E(i) \leq \epsilon, 1 \leq i \leq n} (c_i) \, . \quad (3) \]

Note that, \( E(i) \) is an approximation of the actual error rate of the strings with costs smaller than or equal to \( i \), and the \( \mathcal{T}_c \) value we seek is the largest one where the curve \( E(i) \) reaches \( \epsilon \) (since the curve can decrease at some points, we should choose the last value of \( c \) to maximize the number of accepted strings for a given \( \epsilon \)).

In practice, different test samples will need different rejection thresholds for a given user-defined error rate. Thus, in practical situations, for a given language model, we should have a representative sample in order to compute \( H_e \). That sample can be obtained by manual labeling of a subset of the strings to be processed, and then, we can use \( H_e \) to dynamically estimate the threshold to be applied to any other sample, as explained in this section. Nevertheless, in cases where manual labeling involves at high cost or it cannot be performed for some reason, other techniques addressed to estimate the histogram \( H_e \) of a language model exclusively from known histograms of other language models can be used [2].

Also, different parsing algorithms will give different cost distributions. We have used an Stochastic Error-Correcting Parsing model in the experiments of next section, but the proposed method could be used with other kinds of models if they comply with the requirement that a transformation cost is obtained for each OCR hypothesis.

\section{Experiments}

The experiments have been designed to evaluate the accuracy of an error-rate estimation, comparing it with the real error obtained after applying the post-processing method proposed in [13] to the OCR hypotheses from four handwritten form fields with contents belonging to different language models very typical in commercial forms: names, surnames, towns and provinces. In this case, from Spain.

A training set has been used to compute the histogram \( H_e(c,w) \) for each language model using a window size \( w = 0.5 \), and two test sets from very different batches of industrially OCR-processed forms have been
selected to verify the accuracy of the method: Test 1 from a handwritten form similar to the training set, and Test 2 from a completely different kind of form, with writers extracted from a different socio-geographic population sample. The number of strings of the whole language models and the size of the sets are shown in table 1.

<table>
<thead>
<tr>
<th></th>
<th>Language</th>
<th>Training</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Names</td>
<td>66363</td>
<td>4574</td>
<td>1959</td>
<td>176</td>
</tr>
<tr>
<td>Surnames</td>
<td>97157</td>
<td>9362</td>
<td>4012</td>
<td>319</td>
</tr>
<tr>
<td>Municip.</td>
<td>8200</td>
<td>22371</td>
<td>9848</td>
<td>154</td>
</tr>
<tr>
<td>Provinces</td>
<td>62</td>
<td>13465</td>
<td>5770</td>
<td>175</td>
</tr>
</tbody>
</table>

In figure 2, we can see the approximated error rates computed using equation 2, and the real error rates, as a function of the transformation costs of the strings of the field “Name” (Spanish first names) from Test 1. Real errors at a cost $c$ were computed as the number of strings “erroneously corrected” having a cost up to $c$, divided by the total number of strings having a cost in the same interval. The small differences between both curves show that a good approximation is possible in this case. The peak in the left tail of the real error curve can be explained by the high relative weight that the lowest costs have in the cumulative process.

In figure 3, the deviations between the real error rate and the error rates obtained using the proposed threshold estimation method (Equation 3) for different fields and language models in Test 1 are shown. The results are all in a narrow range around 0, and perfectly useful in a practical application, allowing the user to specify the maximum tolerated error quite effectively.

In figure 4, corresponding to Test 2, we can see that the difference between the real errors and the errors obtained for the different fields are much larger. Test 2 was acquired in very different conditions and corresponds to a very particular geographical region, where the provinces and municipalities are a small subset of the general language models used (almost all of them from Catalonia). The names do not show a significant bias from the general population and the error prediction is kept under reasonable values. The surnames give rise to moderately worse results, but the municipalities and provinces differ considerably given the very small coverage of the sample. As expected, this points out that a representative training set must be used to obtain a reliable error rate estimation.

As a reference, in table 2, the different values for the $T_c$ threshold required in the different cases to keep a real error rate of $\epsilon = 2\%$ are shown.

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Names</td>
<td>4.60</td>
<td>4.84</td>
<td>3.78</td>
</tr>
<tr>
<td>Surnames</td>
<td>3.25</td>
<td>3.27</td>
<td>3.48</td>
</tr>
<tr>
<td>Municipalities</td>
<td>5.39</td>
<td>5.51</td>
<td>5.05</td>
</tr>
<tr>
<td>Provinces</td>
<td>7.48</td>
<td>7.21</td>
<td>4.67</td>
</tr>
</tbody>
</table>
Figure 4. Difference between the real error and the one obtained using the proposed threshold estimation method for different language models in Test 2.

6 Conclusions

A method for the automatic estimation of a threshold on the reliability index, or transformation cost, of strings coming from OCR hypotheses has been proposed and tested with different training and test sets. The automatic estimation of a rejection threshold could allow the user to approximate a pre-defined expected error rate. The results suggest that a good prediction ability can be expected if the training set has a good coverage of the language and the test set is not strongly biased, even if the kind of documents, writing styles and scanning processes are very different.

References


