Rejection Threshold Estimation for an Unknown Language Model in an OCR Task*

Joaquim Arlandis, Juan-Carlos Perez-Cortes,
J. Ramon Navarro-Cerdan, and Rafael Llobet

Instituto Tecnológico de Informática
Universitat Politècnica de València
Camí de Vera s/n, 46071 València, Spain
{arlandis,jcperez,jonacer,rllobet}@iti.upv.es

Abstract. In an OCR post-processing task, a language model is used to find the best transformation of the OCR hypothesis into a string compatible with the language. The cost of this transformation is used as a confidence value to reject the strings that are less likely to be correct, and the error rate of the accepted strings should be strictly controlled by the user. In this work, the expected error rate distribution of an unknown language model is estimated from a training set composed of known language models. This means that after building a new language model, the user should be able to automatically "fix" the expected error rate at an acceptable level instead of having to deal with an arbitrary threshold.

Keywords: Error rate, rejection threshold, language model, error-correcting parsing, OCR post-processing, regression model.

1 Introduction

Optical recognition of printed or handwritten text is often followed by a post-processing phase that can significantly improve the final performance if some constraints are imposed on the contents of the text. The set of constraints can be formally represented as a *language model* (be it a natural language or a subset of a natural language, a closed list of words or expressions, a code following some pattern, etc.). Forms with fields that are filled-in by hand are typical documents where different models can be defined for each field. Frequent field types are "Name", "Age", "Date", "Country", "Street", "Symptoms", "Incident description", "Id. Number", "Phone Number", etc. The language models associated with each of these fields are widely different in many regards (alphabet, size, complexity, perplexity...) and, unlike the OCR classifier for example, that is often kept unchanged for a reasonable amount of time, new language models appear routinely in the normal form-processing large-scale industrial activity.

* Work partially supported by the Spanish MICINN grants TIN2009-14205-C04-02 and Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and by IMPIVA and the E.U. by means of the ERDF in the context of the R+D Program for Technological Institutes of IMPIVA network for 2010 (IMIDIC-2009/204).

E.R. Hancock et al. (Eds.): SSPR & SPR 2010, LNCS 6218, pp. 738–747, 2010.

Very different techniques have been employed to post-process the OCR hypotheses according to a required model (see section 2) and most of them provide or can be easily modified to provide a reliability index (directly related to the *correction confidence* and inversely related to the *transformation cost*).

Applying a threshold to these costs or confidence values allows the system to reject those strings that are less likely to be correct (those involving a high cost or "effort" to convert the OCR hypothesis to a correct output). Usually, the rejected sequences are submitted to a manual data-entry process and therefore the threshold selection has a high impact in the practical performance and economic benefit 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 cost, in terms of time and money, of the human data-entry process.

In this paper, a technique to estimate the expected error rate distribution of a new, unknown, language model, is proposed. That distribution is used to estimate the rejection threshold of a test sample in order to obtain a given expected error rate. Experiments are presented comparing the accuracy of the estimations in different conditions.

The rest of the paper is organized as follows: section 2 contains an overview of the related work. Section 3 describes how the error rate distribution as a function of the transformation costs can be learned, predicted, and used to compute the rejection threshold. In section 4, experiments and results on error rate estimation for different languages are reported, and, finally, the conclusions are presented in section 5.

2 Related Work

Many works on language modeling have been carried out in the field of continuous speech recognition [10]. Although the requirements are very different, many basic techniques used in that discipline can be applied to OCR tasks with little modification. Word and sentence level models typically apply dictionary search methods, n-grams, Hidden Markov Models, Edit Distance-based techniques, and other character or word category transition models. In [6], an excellent survey of approximate string search methods is presented. 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 [9], [18], [15], [12].

In this work, the error-correcting parsing (ECP) technique has been used to post-process the OCR hypotheses is as described in [15]. It consists of building a finite-state machine from a formal grammar, that accepts (or generates with a certain probability) the strings in the lexicon or language sample. When the model is applied to a candidate word the smallest set of transitions that could not be traversed shows which is the most similar string in the model, and the minimal cost of the selected path is provided as a transformation cost of the input. The classical algorithm, widely used, to find the maximum likelihood path

on a Markov model, and to perform ECP, on a regular grammar, is the Viterbi Algorithm, based on the Dynamic Programming paradigm. The extension of the Viterbi algorithm used in this work is described in [1].

The construction of the finite-state machine has been performed using a grammatical inference algorithm that accepts the smallest k-Testable Language in the Strict Sense (k-TS language) [19] consistent with a task-representative language sample. The set of strings accepted by such an automaton is equivalent to the language model obtained using n-grams, for $n{=}k$. The stochastic extension of the basic k-TS language is performed through a maximum likelihood estimation of the probabilities associated to the grammar rules, evaluated according to their frequency of utilization by the input strings. This computation is carried out incrementally and simultaneously with the inference process.

Given the impact of the quality of the confidence estimation on the practical use of an OCR system, many recent works exist that deal with this problem. The work of Landgrebe [13] proposes a modified version of the ROC curve, where a factor to tune the number of expected false positives is introduced in order to tackle with imprecise environments. Other works directly related to post-processing in OCR and text recognition tasks, propose rejection strategies oriented to yield reliable confidence measures [3], [5], [16]. The use of confidence measures has also been specially and traditionally studied in the Speech Recognition and Natural Language Processing areas.

The particular problem of automatic rejection threshold estimation has also applications in economics, medicine, network management, signal processing, and others. A statistical approach often used in many different areas is based on the conventional Monte Carlo techniques, where the thresholds are set according to the distribution percentiles of the measures (or cost functions). These approaches demand very large number of samples to be useful.

Also, statistical methods have been developed, like in [7], where threshold estimation is studied in the context of regression. In sensor systems, where large amounts of data are usually available, the target detection is seriously affected by false positives, and a special effort has been made to improve their behavior. Thus, Ozturk *et al.* [14] used the generalized Pareto distribution to approximate the extreme tail of the distributions of radar measures, and propose the ordered sample least squares method for estimating the parameters of the distributions. 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 propose a way to adaptively maintain low false positive rates and to overcome differences between the model assumptions and the real data.

In other Pattern Recognition tasks, the problem of rejection threshold estimation has also been studied. For instance, in [2], several methods for estimating speaker-independent and speaker-dependent decision thresholds for automatic speaker verification were compared using only relevant parameters estimated from training data.

In handwritten numeral recognition, He *et al* [8] 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 [11] 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.

In the context of many real tasks, specifically estimating an automatic rejection threshold from an user-defined expected error rate would alleviate the problem of dealing with arbitrary (in practice) confidence measures. In this sense, a closer goal to the one presented in this work has been proposed by Serrano *et al* in the context of error supervision in interactive-predictive handwriting recognition [17]. The objective 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.

3 Approach

If we take a representative sample of strings consisting of OCR hypotheses, and compute the transformation costs using a post-processing algorithm (in our

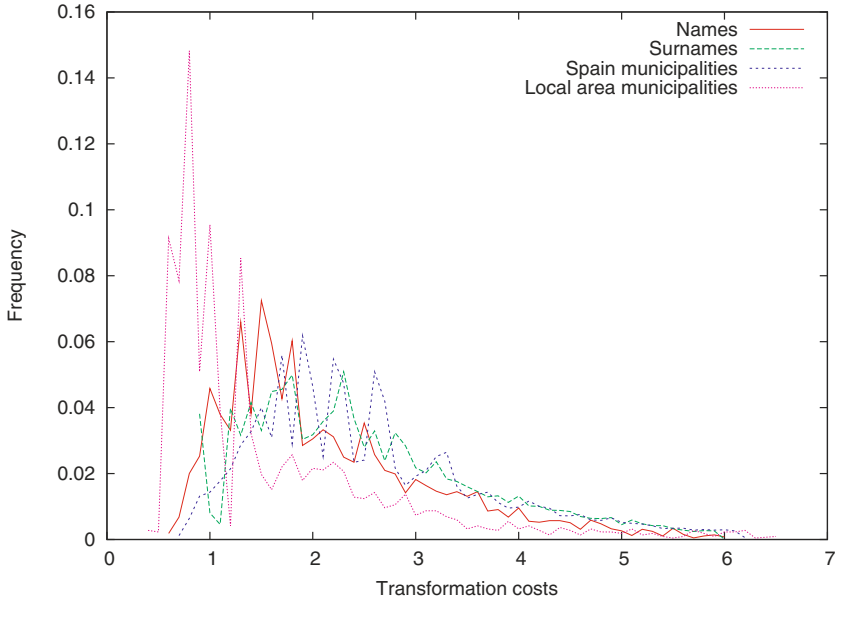

Fig. 1. Histogram of the correction costs of OCR hypotheses strings from four different language models (Spanish names, Spanish surnames, all Spanish towns, and towns from a local region: Comarca de "La Ribera Alta")

case, ECP on a k-TS language [15]), the distribution obtained varies widely for different language models, as can be seen in Figure 1. This means that choosing a consistent rejection threshold is nothing but trivial, since the number of accepted and rejected strings for a given threshold will be very different depending on the characteristics of the language model. Also, moving the threshold value slightly can lead to unpredictable changes on the ratio of accepted/rejected strings.

Therefore, a more predictable confidence index is needed. A technique to estimate the error rate distribution of a test sample as a function of the transformation costs is proposed, consisting on the following steps:

- Given a set of transformation costs obtained from a representative sample of manually labeled OCR hypotheses strings from a language, the error rates associated to each cost *(error rate distribution)* are learned, and then used to find the rejection threshold for new samples of the same language. As described in the next section, the error rate distribution can be used to estimate the rejection threshold for a given expected error rate.
- When a new language model is defined in the system, an automatic way to estimate its error rate distribution that uses exclusively characteristics measured directly on the language model by means of regression techniques is proposed. This way, the time-consuming process of acquisition, OCR and manual validation of a significant amount of strings is avoided. This is specially important if new language models are needed frequently, even if they are subsets or special variants of known models. In section 3.2, the details of the approach are explained.

3.1 Modeling the Error Rate Distribution of a Language Model

Given a language model and a set of transformation costs obtained using a post-processing algorithm with a representative sample of OCR hypothesis strings (for which the ground-truth transcriptions have been manually obtained), a smoothed histogram $H_E(c)$ of error rates for different costs c can be computed using the expression,

$$H_E(c, w) = \frac{|S^-_{c,w}|}{|S_{c,w}|} \tag{1}$$

where w is a smoothing window size parameter, $|S^-_c|$ is the number of strings "erroneously corrected" into an incorrect string having a cost between $c - w$ and $c + w$, and $|S_c|$ is the total 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 c instead of a fixed cost interval. In Figure 2, a histogram H_E obtained using the post-processing algorithm of [15] on different language models is shown.

We can easily find the rejection threshold $\mathcal{T}_c$ required to obtain a given error rate ϵ on a test sample S' by accumulating averaged values of H_E according to increasing values of c,

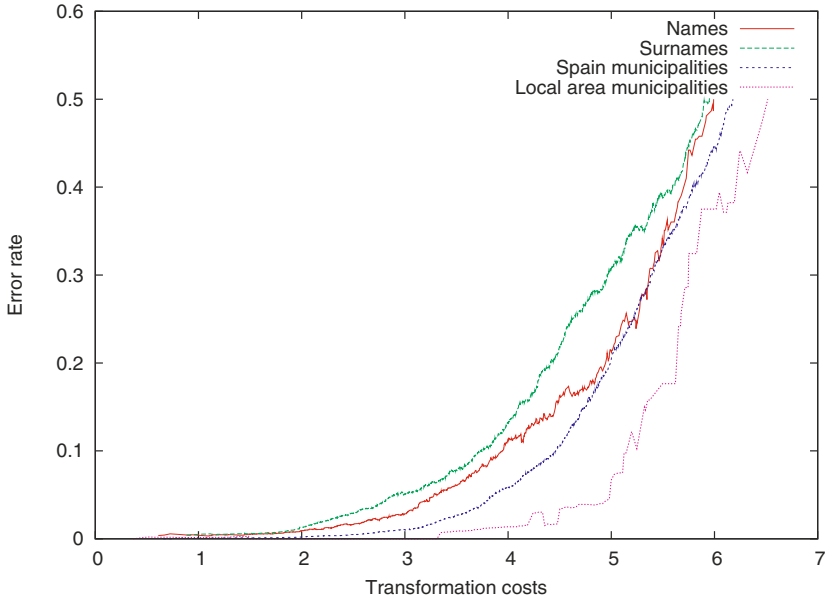

Fig. 2. Error rate histogram H_E, for the sample of language models plotted in Figure 1 using $w = 0.5$ (Equation 1)

$$E(i) = \sum_{c=c_1}^{c_i} \frac{H_e(c,w)}{i} \ , \ c \in S'$$

where the value of $E(i)$ at each point is the average error rate of the strings with costs smaller or equal than c_i. Then, the $\mathcal{T}_c$ value we seek is the largest one where the curve reaches ϵ (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 ϵ).

E can be seen as a cumulative averaged version of H_E for a given test sample and it can be used to approximate the appropriate cost threshold to use when we want to fix the expected error rate. In practice, different test samples will require different rejection thresholds for a given user-defined error rate.

3.2 Estimating the Error Rate Distribution of New Language Models

Let H_C be the histogram of the transformation costs of the strings that belong to a language (positive sample). H_C can be easily obtained from the list of positive strings because it does not depend on the OCR process. Figure 3 shows the histogram H_C of the same four language models shown in Figure 1.

Both figures 1 and 3 suggest that there is a correlation between the distributions of the costs of OCR hypothesis strings (many of them having errors), and

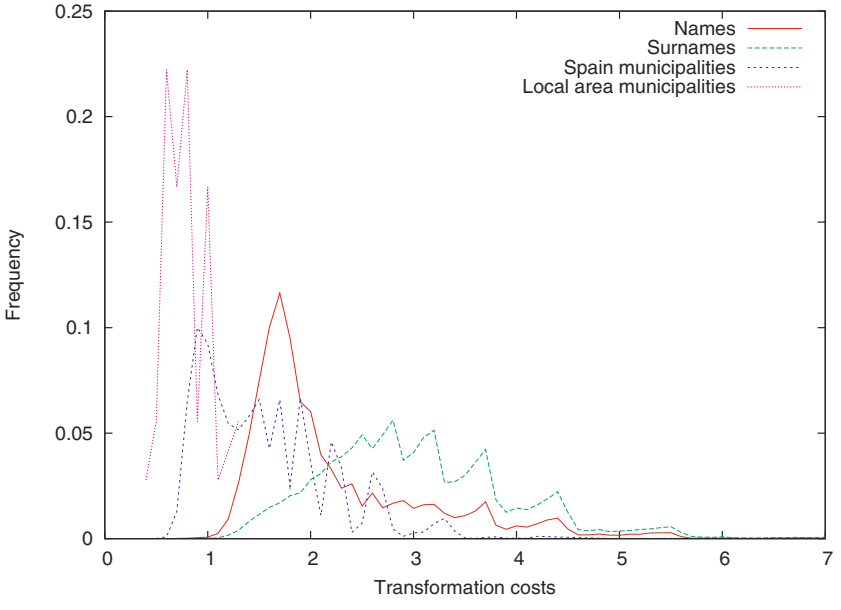

Fig. 3. Histogram H_C of the correction costs of strings belonging to four different language models (Spanish names, Spanish surnames, all Spanish towns, and towns from a smaller local region: Comarca de "La Ribera Alta")

positive samples of the same language model (without errors). And, as already mentioned, the cost distributions of different languages clearly differ.

Assuming the above statement, we propose that a training set composed by features extracted from the histograms H_C and H_E of a set of known languages models is used to build a regression model able to predict the expected error rate distribution $\widehat{H_E}$ (target output) of a new language based on features extracted from its H_C histogram (inputs).

Several regression methods have been tested. The results obtained and the details on these methods and their parametrization are described in the next section.

4 Experiments

The goal of the experiments has been to measure the capability of the regression techniques to learn a function that approximates the error rate distribution $\widehat{H_E}$ of new language models from a model built using features extracted from known language models as described in the former section.

The four different language models shown in the figures of the previous sections have been used to perform a leaving-one-out estimation. They are the names and surnames in the last census of Spain: 66363 names and 97157 surnames with probabilities derived from their frequencies in the census, all Spanish municipalities (8201 towns without frequencies, and 35 municipalities, without frequencies,

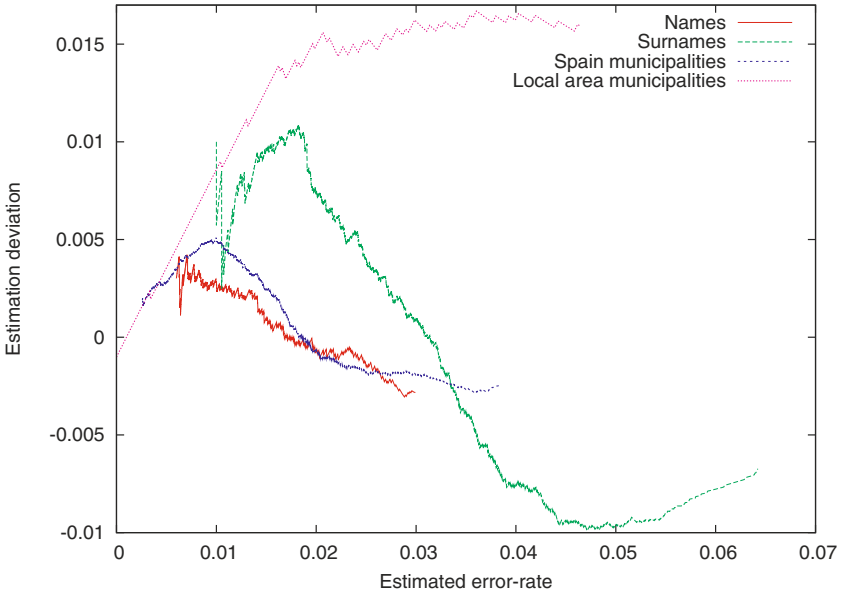

Fig. 4. Differences between the estimated error rate and the real error rates for the four language models in a leaving-one-out experiment

from a local region: Comarca de "La Ribera Alta"). These languages have been chosen since they are representative of real tasks and span a wide range of sizes.

For each experiment, a single regression model has been built using 2000 OCR hypothesis strings chosen randomly from each language model. To train the model, a number of features of each language including transformation cost and error rate to describe the distributions of H_E, and statistics like mean, median variance, percentiles, coefficient of variation and frequencies of the bins describing the distribution of H_C have been combined. The target output variable for the regression is the error rate, measured applying the language model, for each cost.

Several regression models have been tested (Support Vector Machines for Regression, Radial Basis Functions and a Multilayer Perceptron, with similar results). The results are provided in terms of estimation deviation, i.e., the difference between the estimated error –computed as explained in section 3.1, but on the error rate distribution $\widehat{H_E}$ estimated by the regression model– and the real error measured in the test set.

In Figure 4, the estimation deviation is plotted against the estimated error, for the four language models. For the test of each language model, the regression model has been built using the other three language models.

We can see that the estimation can be useful in all cases, but it is more accurate in the case of Names and Spain Municipalities. In practice, the typically acceptable error rates are in the range of 1% or 2% (between 0.01 and 0.02 in

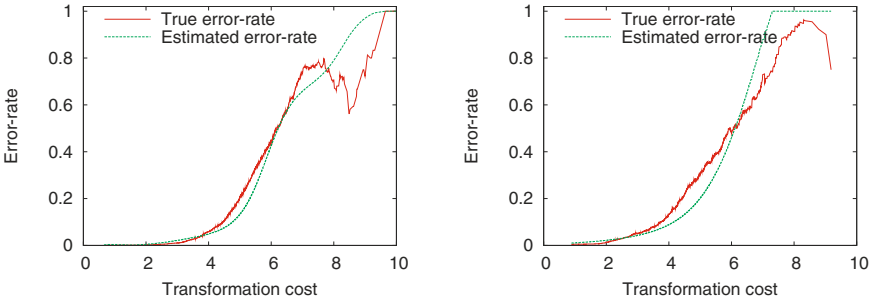

Fig. 5. Error rate histograms, H_E, and estimated error rate histograms $\widehat{H_E}$, for the language models of Spain municipalities (left) and Spanish surnames (right)

the figures). In that useful range, the error deviations are small enough to be directly usable in the two best languages, and a good starting point for a slight empirical adjustment in the case of the two worst languages. With a larger set of language models to train the regression model, we think these results can be significantly improved.

In Figure 5, the error rate histogram of the test sample, H_E, along with the estimated error rate histogram, $\widehat{H_E}$, are plotted for two of the language models studied.

5 Conclusions

We have presented a method for the estimation of the expected error rate distribution of an unknown language model, so that a user can establish the error rate at an acceptable level and the system estimates the rejection threshold automatically.

Experiments where a regression model is built using OCR hypotheses from a set of known languages have been performed, and the model is tested against a new language. The results show a useful behavior, with reasonably accurate estimations of the rejection threshold in the typically practical range of error rates. As a future work, we plan to train the regression model with a larger set of language models.

References

1. Amengual, J., Vidal, E.: Efficient error-correcting viterbi parsing. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(10), 1109 (1998)
2. Lindberg, J., Koolwaaij, J., Hutter, H., Genoud, D., Pierrot, J., Blomberg, M., Bimbot, F.: Techniques for a priori decision threshold estimation in speaker verification. In: Proceedings RLA2C, pp. 89–92 (1998)
3. Bertolami, R., Zimmermann, M., Bunke, H.: Rejection strategies for offline handwritten text line recognition. Pattern Recognition Letters 27(16), 2005–2012 (2006)

4. Broadwater, J., Chellappa, R.: Adaptive threshold estimation via extreme value theory. IEEE Transactions on Signal Processing 58, 490–500 (2010)
5. Gandrabur, S., Foster, G.F., Lapalme, G.: Confidence estimation for nlp applications. TSLP 3(3), 1–29 (2006)
6. Hall, P., Dowling, G.: Approximate string matching. ACM Surveys 12(4), 381–402 (1980)
7. Hansen, B.E.: Sample splitting and threshold estimation. Econometrica 68(3), 575–604 (2000)
8. He, C.L., Lam, L., Suen, C.Y.: A novel rejection measurement in handwritten numeral recognition based on linear discriminant analysis. In: 10th Intl. Conf. on Document Analysis and Recognition, pp. 451–455. IEEE Computer Society, Los Alamitos (2009)
9. Hull, J., Srihari, S.: Experiments in text recognition with binary n-gram and viterbi algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence 4(5), 520–530 (1982)
10. Jelinek, F.: Up from trigrams, the strugle for improved language models. In: European Conf. on Speech Communication and Technology, Berlin, pp. 1037–1040 (1993)
11. Kae, A., Huang, G.B., Learned-Miller, E.G.: Bounding the probability of error for high precision recognition. CoRR, abs/0907.0418 (2009)
12. Kolak, O., Resnik, P.: Ocr post-processing for low density languages. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT), pp. 867–874. Association for Computational Linguistics (2005)
13. Landgrebe, T., Paclík, P., Duin, R.P.W.: Precision-recall operating characteristic (p-roc) curves in imprecise environments. In: International Conference on Pattern Recognition ICPR (4), pp. 123–127 (2006)
14. Ozturk, A., Chakravarthi, P.R., Weiner, D.D.: On determining the radar threshold for non-gaussian processes from experimental data. IEEE Transactions on Information Theory 42(4), 1310–1316 (1996)
15. Perez-Cortes, J., Amengual, J., Arlandis, J., Llobet, R.: Stochastic error correcting parsing for ocr post-processing. In: International Conference on Pattern Recognition ICPR-2000, Barcelona, Spain, vol. 4, pp. 405–408 (2000)
16. Pitrelli, J.F., Subrahmonia, J., Perrone, M.P.: Confidence modeling for handwriting recognition: algorithms and applications. International Journal of Document Analysis 8(1), 35–46 (2006)
17. Serrano, N., Sanchis, A., Juan, A.: Balancing error and supervision effort in interactive-predictive handwriting recognition. In: International Conference on Intelligent User Interfaces (ICIUI), Hong-Kong, China (2010)
18. Tong, X., Evans, D.A.: A statistical approach to automatic ocr error correction in context. In: Fourth Workshop on Very Large Corpora, pp. 88–100 (1996)
19. Garcia, P., Vidal, E.: Inference of K-Testable Languages in the Strict Sense and Application to Syntactic Pattern Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 12(9), 920–925 (1990)