

A Statistical Correction-Rejection Strategy for OCR Outputs in Persian Personal Information Forms

R. Mehran, A. Shali, and F. Razzazi

Abstract--In this paper, a MAP statistical modeling approach has been utilized to correct and verify Persian names and surname OCR outputs. In addition, an efficient Neural Network based rejection method has been presented and tested. Due to large variety of Persian surnames, a statistical grammar has been added to the MAP strategy, to make new surnames, which are not included in the dictionary. The model has been analytically formulated and practically implemented. The achieved results show a large character and word error reduction while the calculation increase is negligible in comparison with character recognition complexity.

Index Terms-- Natural Language Processing, OCR, MAP, Neural Networks, Persian language.

I. INTRODUCTION

Automatic Document Analysis has a key role in office automation and data entry systems. Meanwhile, personal information occupies the essential part of almost every handwritten questionnaire. At the heart of every document processing system, there is a natural language processing unit with the ability of making balance between human language and the primary output of the system. Even in the best OCR systems, there are subtle errors due to misconceiving the characters by the system [1].

Fortunately, there are a limited number of target words in a data entry system for personal information. This fact motivates NLP researchers to develop algorithms to correct the OCR output by using language information of a bounded set of target words. There is a rich literature on embedding language models into OCR data entry systems for English language [2]-[4]. However, neither the OCR systems, nor the appropriate language models have been developed enough in Persian language [5].

Last year, we have developed a Persian OCR data entry system for personal information forms [1]. To correct the character recognition output of the system by linguistic information, we need to develop a language model for our target sets (e.g. first names, surnames, locations, etc...). [4], [6] In this paper, we will present our approach in developing Persian language models for first name and surname target sets and using this model in a MAP [7] strategy to correct the OCR outputs. In addition, we will

present a rejection method based on a neural network classifier.

The target sets of both girls and boys names are limited enough to model them statistically whereas the process of finding the best choice for the surname field in Persian handwritten forms is not so effortless. The structure of almost every Persian surname consists of a composition of words, one or more prefixes, and one or more suffixes. Each subcomponent could nearly be any Persian noun. As a sample, in a database with the population of 170,000 from Iranian student names with only 4900 distinct first names (2200 for boys and 2700 for girls), there are more than 70,000 nonidentical surnames. This complex feature, which is imposed by the language, has been dictated to every NLP algorithm that deals with Persian surnames.

In this paper, a lexicon-based approach for processing Persian surname has been proposed and evaluated. By combining the lexicon-based approach with statistical grammar strategies, we introduced a novel solution to the proposed problem. Finally, a rejection management unit follows this spell correction unit, providing a safeguard for the actual implemented system. The process of the rejection unit is considered as a classification task of confirming or rejecting the overall output of the character recognition and natural language processes. Hence, the rejection unit is chosen to be a MLP neural network. As the result shows, this structure has considerably improved the ratio of recognition.

II. SYSTEM STRUCTURE

The basic idea of the proposed statistical NLP system is based on modeling the OCR system as a noisy communicational channel. The behavior of the OCR system has been estimated before developing the NLP unit. Firstly, the probability distribution of Persian used surnames was assumed to be known. Then we have applied the MAP strategy to recover the errors of our handwritten Persian OCR system. The Fig 1. Shows the communicational channel model for NLP and the MAP strategy argument is mentioned in (1), where W_i is the input word, W_r is the recognized word, W_i^* is the corrected word by NLP and P is the corresponding probability value.

$$W_i^* = \underset{w_i}{\operatorname{argmax}} P(w_i | w_r) = \underset{w_i}{\operatorname{argmax}} P(w_r | w_i) \cdot P(w_i) \quad (1)$$

This work was sponsored by PayaSoft Co. and Isfahan Science and Technology Town (ISTT).

R. Mehran, A. Shali and F. Razzazi are with PayaSoft Co., Tehran, Iran (e-mail: info@payasoft.com Tel: (+9821) 8305430, Fax: (+9821) 8305430)

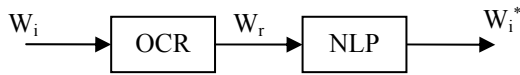


Fig.1 - Communication Channel model for NLP and the MAP strategy.

The knowledge of OCR system behavior enables us to build a statistical model of the system. This model has been smoothed by Good-Turing method in order to get a better estimation of conversion probabilities of characters [7], [8]. The implemented model was a mono-gram conversion model of alphabetic characters.

The complexity of Persian surnames is based on its composite structure and since the database of all Persian surnames is not available, the result of implementing standard MAP strategy would not be satisfactory [9]. However, other fields of the questionnaire have been corrected using MAP.

The main idea of our approach is to build a lexicon based NLP for Persian surnames, which by combining the lexical parts, it could create new acceptable surname in accordance with OCR character outputs. This new strategy in processing surnames is also based on a statistical model of Persian language and its lexical parts. The process of building surnames could yield new words, which they are not available in the database; however, they would most likely be the same character string as the handwritten surname. In the lack of a comprehensive database for Persian surnames and in the absence of a useful lexicon in Persian, we decided to build our own surname database and lexicon [1].

A thorough method of parsing for every character string is applied in which it covers every possible subcomponent of the full-length string. In the parsing process, every word or subcomponent is parsed into two parts. The first part is always conceptually considered as a "prefix" and the last part is considered as a "suffix". Then each of these subcomponents is parsed until reaching a predefined minimum length. After this process, every possible combination of these components is checked and a cost value is assigned for each combination by MAP strategy.

By considering the probability of every subcomponent of the word in the lexicon and the error probability of its containing characters in the OCR model, a cost value is assigned to it. We save the ten-best answers for each subcomponent and then by combining these subcomponents we produce a set of combinational surnames. This means, if there are n valid parts in a surname we have produced a set of surnames with 10^n members. In the combination procedure, the cost value of each specific composition is computed by a bi-gram model.

In the estimation phase, the necessary lexicon for cost evaluation process has been built based on manual database prefix-suffix labeling. In addition, all required bi-gram conversion probabilities are computed.

In correction phase, the cost value of every combination of prefixes and suffixes can be computed as (2) where C stands for cost and P is the probability value.

$$C(prefix, suffix) = \log(1 / P(suffix | prefix)) \quad (2)$$

As an example, Fig. 2 shows an equivalent combination of words that might compose a surname in Persian if they are translated and positioned correctly in English. The depicted parsing procedure is based on the mentioned prefix and suffix assumption of every Persian word.

III. REJECTION METHOD

Although there is a rich literature on rejection management units for pattern recognition, there is few cases of applying the rejection process after the natural language processing unit in the field of text processing [10], [11]. In our approach, we have utilized a three-layered MLP classifier as the rejection unit with ten inputs in the first layer, 40 neurons in the hidden layer and one output as the rejection flag. These ten neurons receive the ten best cost-probability values of first name or surname. Hence, the classifier decides the rejection or confirmation of the best output of the NLP process by costs of 10 best outputs in each case of first names or surnames. The neural network was trained on a set of OCR system outputs which, those set of data were manually confirmed. By utilizing this rejection method on the test sets, the results show a significant shift in the recognition rate has been achieved. Meanwhile it ensures a low rate of rejection and very small ratio of number of misrejection cases to the number of total rejection cases. By applying this method, the characteristics of the actual system, which impose the necessity of high rate of accuracy and efficiency, was achieved.

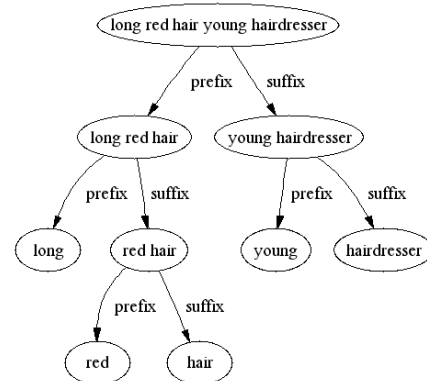


Fig.2 – Example of parsing procedure diagram.

IV. RESULTS AND DISCUSSION

We had tested our NLP approach on an OCR system with a hierarchical neural network structure with the average of 81% accuracy on character recognition. Initially the system was designed by the assumption of closed vocabulary while on the contrary the exact vocabulary of the Persian surnames was not available. As the Table I and II show, composing new words out of some principle word components (prefix and suffix), provides considerable results for surname field. In the rejection phase, the results show that there is valuable

information embedded in the relations within the cost values assigned during NLP process. Neural network classifier processes this information fairly well which it would not be achieved by simple comparison of cost values. The proposed rejection strategy could be performed on each field of a personal information form, only by training the neural network classifier with a proper training set. This system was tested on nearly 160,000 personal information forms and the results of the early 6,109 forms are presented in Table II. The more cases to be presented to the system the more accurate statistical model is being built and the results on the whole set of forms show a slight improvement to the results which are presented in the Table II.

The proposed NLP process was carried out on a character stream, which was constructed based on the best choice of OCR system for each character. As the probability of each character in the specified forms in Persian language was computed during the implementation, the next step for improving the results would be adding a dynamic programming search to find the optimal character stream, prior to the proposed NLP process. This could be achieved by using a Viterbi search on an $n \times m$ matrix of n-best OCR outputs, where m is the length of the string and n is the number of best choices for each character. For further improvement, we are considering to build a measure of confidence based on the OCR output posteriori probabilities. This will be used as an additional data for the rejection criteria. Finally, we would suggest more complex statistical grammar for Persian surname to be used and more accurate conversion probabilities to be computed in order to achieve better results.

Table I – Correction results for surname field

	With parsing	Without parsing
First output	71.86 %	62.43 %
10-best output	84.02 %	68.81 %

Table II – Rejection results on 6109 personal information forms

	Name	Surname
Recognition Rate Without Rejection	89.04 %	77.78 %
Recognition Rate After Rejection	92.23 %	90.85 %
Ratio of Rejection to Total Number of cases	5.18 %	20.3 %
Ratio of Wrong Rejections to total number of Rejections	30.91 %	26.62 %

ACKNOWLEDGMENT

The authors wish to express their special thanks for supports and collaborations of Iran's National Organization for Development of Exceptional Talents (NODET).

REFERENCES

- [1] F. Razzazi, R. Mehran, A. Shali, M. Soleymani, M. Afshar, H.Pirsiavash, "OCR project technical report: Isolated Handwritten Persian optical character recognition system with bounded goal set", *Isfahan Science & Technology Town*, 2003, (In Persian).
- [2] R. Rosenfeld, "Two decades of Statistical Language Modeling: Where Do We Go From Here?", *Proceedings of the IEEE*, 2000, pp. 88-96.
- [3] E. Brill, R. Florian, J. C. Henderson, L. Mangu, "Beyond N-Grams: Can Linguistic Sophistication Improve Language Modeling?", *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, Vol. I, 1998, pp. 186-190.
- [4] M. Shridhar, F. Kimura, B. Truijen, G.F. Houle, "Impact of lexicon completeness on city name recognition", *Proceedings of IWFHR-8*, 2002, pp. 513-518.
- [5] F. Razzazi, M. Soleymani, H. Pirsiavash, "Design and implementation of a system for recognition of Isolated handwritten Persian characters with a known dictionary of words: a Hierarchical neural network approach", *First Iranian National Conference on Machine Vision*, Birjand, Iran, 2001, (In Persian).
- [6] S.F. Chen, "Building Probabilistic models for natural Language" PhD thesis, Harvard University, Cambridge, Massachusetts, Cambridge, Mass., 1996
- [7] M.H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Co., New York, NY, 1970.
- [8] S. Chen, J. Goodman, "An empirical study of smoothing techniques for language modeling", *Proceedings of the 34th Meeting of the Association for Computational Linguistics*, 1996, pp. 310-313.
- [9] W. A. Gale "Good-Turing Smoothing without Tears", *Journal of Quantitative Linguistics*, 1995, pp. 217-254.
- [10] A. Fatholahzadeh, M.J. Nategh, "Lexicon Base in Persian for Object Manipulation", *Proceedings of PMEC*, Amir Kabir University of Technology, Tehran, 1993, pp. 52-61.
- [11] CK Chow., "On optimum recognition error and reject tradeoff", *IEEE. Trans. Inform. Theory*, IT-16, 1970, pp.41-46.
- [12] J. Arlandis, J.C. Perez-Cortes, and J. Cano., "Rejection strategies and confidence measures for a k-nn classifier in an ocr task", *16th. International Conference on Pattern Recognition ICPR-2002*, volume 1, 2002, pp. 576-579.