

# A Lexicon Reduction Strategy in the Context of Handwritten Medical Forms

Robert Milewski

Srirangaraj Setlur

Venu Govindaraju

Center of Excellence for Document Analysis and Recognition  
University at Buffalo, State University of New York  
Buffalo NY 14260

{milewski, setlur, govind}@cedar.buffalo.edu

## Abstract

Traditional handwriting recognition algorithms rely heavily on small lexicons and clean word images. Unfortunately, emergency medical documents do not satisfy either of these conditions. This is a significant road-block that is hampering efforts to rapidly convert valuable offline healthcare handwriting data into digital content that can be efficiently mined for information. This paper describes a strategy whereby given an image representing a noisy handwritten word from a medical document, and a large lexicon consisting of English, medical and pharmacological words, symbols, abbreviations and acronyms, significantly reduces the size of the lexicon while keeping the unknown desired entry within the lexicon. The approach combines geometric interpretations of the word image along with contextual inference of concepts to reduce lexicons for word recognition. The data extracted can then be efficiently and securely disseminated for epidemiological and outbreak detection/analysis. Experimental results on NY State PCR forms are reported.

## 1. Introduction

In the United States, any pre-hospital emergency medical care provided has to be documented rigorously. Departments of Health of each State prescribe a standard medical form to be used in documenting all information on the patient's status and treatment from the moment the rescue effort begins until the patient is transported to the Hospital. State laws require emergency personnel to completely fill out this form for each patient.

Data for this research, in the form of live research copies of the New York State Pre-Hospital Care Report (PCR)[1] (see Figure 1), has been obtained under an agreement with the Western Regional

Emergency Medical Services (WREMS) division of the New York State (NYS) Department of Health. Each PCR is stored as a 300 DPI color JPG image. Computations are only performed on fields containing the relevant medical information; more specifically, no computations are performed on fields containing patient identifying information or on PCRs involving patients with behavioral disorders. The PCR is a form used to gather vital patient information that is used by health care administrators as a resource to identify trends through macro analysis. Currently, PCRs are mostly paper forms and the process of keying this data into a database that can be processed and mined for trend information can take up to several years in many states. A nationwide database of PCR data would be invaluable for a public health syndromic surveillance system.

Figure 1 New York State Pre-Hospital Care Report [1]

Section 2 provides an overview of the proposed technique. Section 3 illustrates the procedure for

training the artificial neural network (ANN). Section 4 describes the design of the recognition algorithm and the lexicon reduction strategy. Performance numbers are reported in Section 5. Section 6 provides a summary and lists future research directions.

## 2. Overview

There are five major zones on the PCR containing the handwritten information of interest: Chief Complaint, Subject Assessment, Objective Physical Assessment, Comments, and Past Medical History. These handwriting areas contain numbers (e.g. 84), symbols (e.g.  $\uparrow$  = increase), abbreviations (e.g. ABD = abdominal), acronyms (e.g. CHF = Congestive Heart Failure), anatomical descriptions (e.g. thoracic), medical conditions (e.g. pneumothorax), pharmacological words (e.g. codeine) and common English. The handwritten zones can contain data from a large heterogeneous lexicon and the text often does not fit perfectly within form boundaries. The text zones therefore present a highly challenging recognition problem. While other work has been performed in the area of unconstrained handwriting, it has been limited to a large lexicon in the isolated domain of common English [2].

The form available for processing and data mining is a carbon copy. The carbon mesh residue in various locations on the form, and broken/unnatural handwriting due to ambulance movement and emergency environments, add further complexity to the document. There has been little prior work on poorly written words [12]. The rest of the form consists of form elements such as checkboxes, segmented character locations, and segmented digit-only locations. These are less of a challenge from a recognition perspective.

A compact lexicon is constructed before sending each segmented word from the PCR to a lexicon driven handwriting recognition algorithm. Research has shown that the smaller the lexicon the greater the odds of successful recognition, provided that the target word is in the lexicon [3][6][7][9][11]. That is the motivation for the reduced lexicon strategy.

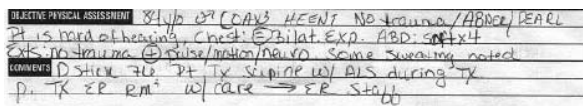


Figure 2 Example of PCR Handwriting to be Recognized

An enhanced binarized representation is produced for each segmented word, which is input to a lexicon

post-processing word recognizer (LPWR); this recognizer uses the lexicon only after all character segmentation and recognition have been performed on the input image. The characters with the highest confidence, determined by the recognizer, are then converted to a list of substring text patterns, and then to several ANN compatible input layer configurations. These configurations are then trained against a concept. This research uses the part(s) of the patients' body, in which the injury/ailment exist as signs (something which can be seen by the healthcare professionals) or symptoms (something felt only by the patient), as the concept. The concept is determined by a healthcare provider, with experience in emergency civilian ambulance rescue, for each PCR in the training set. The training overview can be found in Figure 3.

When presented with a new group of three word images, the LPWR is used to produce the substring configurations, which are used to query the ANN for the concept. The words from the lexicon associated with that concept are finally extracted. This pruned lexicon is then submitted to a lexicon driven word recognizer (LDWR) to produce the final recognition results. Results are also shown using the LPWR instead of the LDWR in the final stage. The recognition overview can be found in Figure 3. The analysis and performance of LPWR and LDWR can be found in [8] and [10]. This work builds on prior research in hybridized systems using lexicon driven handwriting recognition in conjunction with artificial neural networks [5].

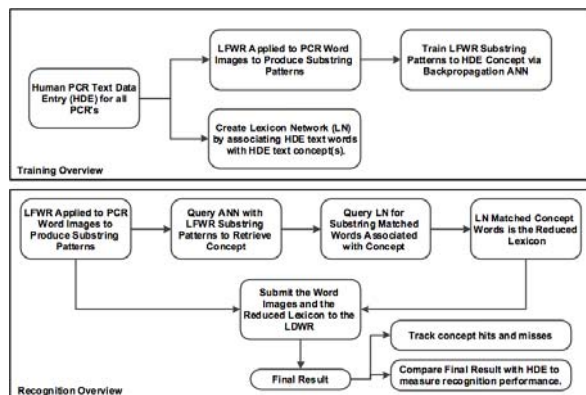


Figure 3 Training and Recognition Procedure Overview

## 3. ANN Training

Figure 4 displays the procedure for training a back-propagation ANN with the bi-gram sequence concepts generated from the LPWR. The ANN is used to associate ailment/injury locations on the patient's body

(i.e. the concept), with each grouping of three adjacent words, on the corresponding medical form. The word grouping is denoted as a phrase and the patient's ailment/injury location is denoted as the target concept to be learnt. Each concept corresponds to one of these 7 unique nodes in the ANN output layer:

- ARMS
- BACK
- CIRCULATORY/CARDIOVASCULAR SYSTEM
- EXCRETORY SYSTEM
- HEAD
- NERVOUS SYSTEM
- RESPIRATORY SYSTEM

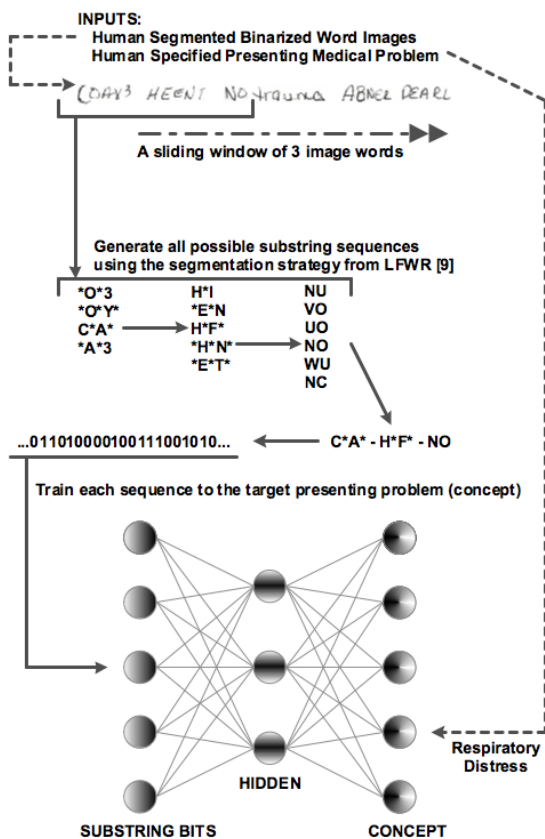


Figure 4 ANN Training Flow

While a greater level of anatomical description is possible, it is unnecessary for classifying information in the pre-hospital emergency scenario. These categories represent all of the patient's injury/ailment locations contained in the PCRs available for this research.

The five handwriting zones from each PCR image are prepared for the ANN input layer as follows: The database is queried for the coordinates of each word (previously segmented by a human).

Binarization, noise removal, and opening operations are applied to each extracted word image. The word images are passed to a lexicon free word recognizer. All possible bi-gram patterns using the highest confidence character score combinations, are produced for each word image. Each pattern is in the format (\*C<sub>1</sub>\*C<sub>2</sub>\*) where C<sub>1</sub> and C<sub>2</sub> are characters whose recognition confidence is above an empirically determined threshold and an asterisk (\*) indicates other character(s) that may or may not be present in that slot. The use of bi-grams over a different n-gram was motivated by these reasons: (1) a significant amount of words used in the PCR (for all PCR's) are of string length two; (2) using an n-gram such that n>=3 would result in words with string length less than n being unmatched.

A Bi-gram sequence, which feed the ANN input layer, are trained to the desired output layer node corresponding to the target concept. The ANN is trained until a stopping condition is met (e.g. either by an epoch metric or by determining that the ANN error, over the complete training set, is within an acceptable threshold).

#### 4. ANN Recognition

Figure 5 displays the procedure for using the LPWR generated bi-gram sequences to determine the concept. The concept and the substring sequences are used to prune the original lexicon to a more acceptable size for use with a LDWR.

Word recognition has 5 phases as illustrated in Figure 5: The individual word images are run through a lexicon free word recognizer. A list of substrings which represent each word is produced. The ANN is queried with all possible bi-gram phrase arrangements for matching concept(s). The concepts and their associated bi-gram sequences are used to query a concept organized lexicon network for a reduced lexicon. The original word images are passed through a LDWR along with the reduced lexicon to determine the recognized word.

#### 5. Results

The test set is organized to have 2 PCR images for each of 7 concepts which comprise the 14 test set images. Those 7 concepts each have between 3 and 15 PCR training examples from the set of 59.

The recognition experiments were performed by supplying the complete lexicon and each word image, from all test images, to the LPWR, LDWR, LPWR+LR+LPWR and LPWR+LR+LDWR

algorithms. For each algorithm, the average recognition rate and the best case recognition rate were determined. The complete lexicon denotes all words from the PCR being recognized in conjunction with those words from the lexicon reduction step.

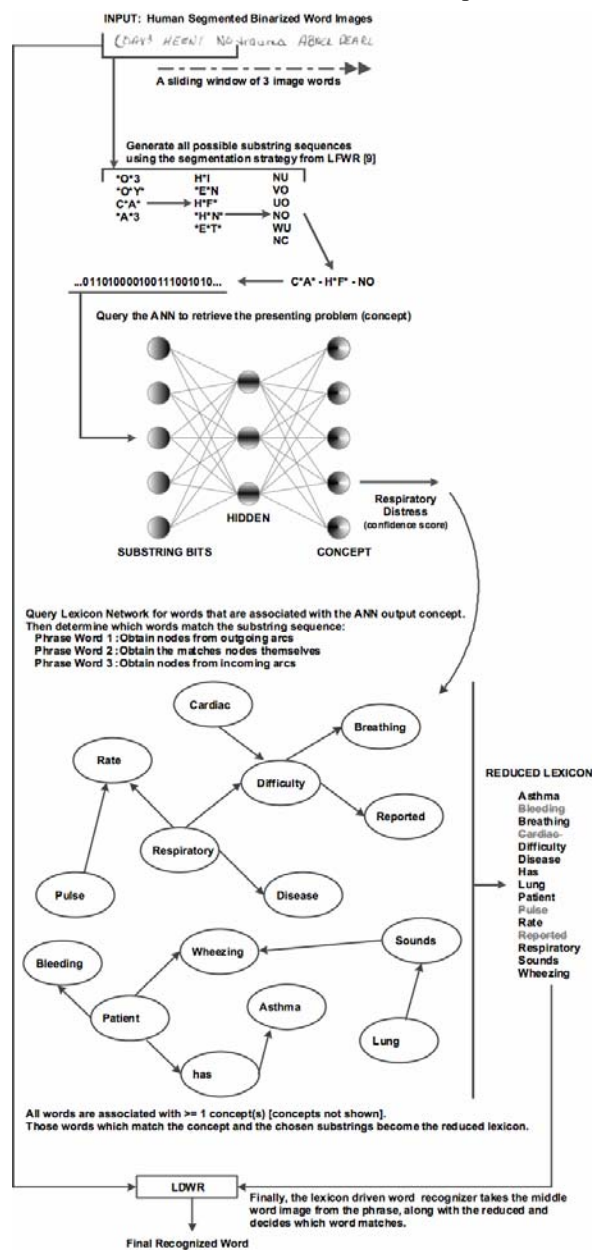


Figure 5 Recognition Flow Example

The performance summary for the test set is provided in Tables 1 and 2; this shows improvement over the LPWR and LDWR performance in prior recognition research on the PCR [4].

TABLE 1: Performance for each PCR

PC R	A	B	C
1	1.6	4.7	4.7
2	27.5	62.5	40.0
3	5.3	16.0	13.3
4	53.6	53.6	58.9
5	23.33	40.0	46.7
6	25.0	29.5	43.2
7	42.2	40.0	57.7
8	25.9	50.0	29.3
9	31.0	24.1	44.8
10	15.1	17.0	26.4
11	35.3	47.1	44.1
12	51.2	60.5	58.1
13	31.6	31.6	45.6
14	10.0	10.0	20.0

Each column corresponds to the performance for the three algorithms: (A) LDWR (B) LPWR+LR+LPWR; and (C) LPWR+LR+LDWR. The recognition value is the percentage of words correctly recognized on the respective PCR image.

TABLE 2: Performance Summary

Training PCR Quantity:	59
Testing PCR Quantity:	14
Lexicon Size:	1,135
ANN Input Layer Feeds:	438,240
LPWR Average:	5.3 %
LDWR Average:	27.0 %
LPWR+LR+LPWR Average:	34.8 %
LPWR+LR+LDWR Average:	38.1 %
LPWR Average Difference:	+29.5 %
LDWR Average Difference:	+11.1 %
LPWR Best:	12.5 %
LDWR Best:	53.6 %
LPWR+LR+LPWR Best:	62.5 %
LPWR+LR+LDWR Best:	58.9 %
LPWR Best Difference:	+50.0 %
LDWR Best Difference:	+5.3 %

Average Lexicon Reduction: R/O 85.4 %  
 Average Reduced Lexicon Size: 1,135 to 166

*Lexicon Size:* The quantity of unique text words (determined by human data entry) across all PCR's.

*ANN Input Layer Feeds:* All possible bi-gram substring sequences generated from the total quantity of PCR words to recognize.

*LPWR [LDWR] Average/Best:* The average [best] recognition performance by the LPWR [LDWR] using an input image word, and the complete lexicon, with no lexicon reduction step applied.

*LPWR+LR+LPWR Average/Best:* The average [best] recognition performance using the presented lexicon reduction strategy with the exception that the LPWR is used with the reduced lexicon, instead of the LDWR at the final recognition stage.

*LPWR+LR+LDWR Average/Best:* The average [best] recognition performance using the presented lexicon reduction strategy.

*LPWR Average/Best Difference:* The average [best] case recognition difference between the LPWR and LPWR+LR+LPWR strategies.

*LDWR Average/Best Difference:* The average [best] case recognition difference between the LDWR and LPWR+LR+LDWR strategies.

*Average Lexicon Reduction:* This percentage measures the quantity of words ruled out (R/O) of the lexicon (i.e. pruning/reducing the lexicon) on average.

*Average Reduced Lexicon Size:* The size of the reduced lexicon (i.e. the quantity of words used as input to the LDWR after the lexicon reduction stage) on average.

## 6. Summary and Future Work

This research has shown that a lexicon reduction of 85.4 % can be achieved while improving the recognition of LPWR+LR+LDWR by 11.1 % on average as well as LPWR+LR+LPWR by 29.5 % on average. This has been made possible by using words which are associated with the ailment/injury location (i.e. the concept) of a patient. This operates under the assumption that all words from the truthed PCR are in the lexicon.

As the time consuming data collection process continues, it will be interesting to see how a larger lexicon and more concepts impact recognition

performance. Additional medical form data may advance the research by allowing various n-gram sizes and the use of multiple concepts in the lexicon reduction stage as well.

## 7. References

- [1] Western Regional Emergency Medical Services. Bureau of Emergency Medical Services. New York State (NYS) Department of Health (DoH). Prehospital Care Report v4.
- [2] Vinciarelli, A., Bengio, S., and Bunke, H. Offline Recognition of Unconstrained Handwritten Texts using HMMs and Statistical Language Models. IEEE Trans PAMI, (2004).
- [3] Carbonnel, S. and Anquetil, E. Lexicon Organization and String Edit Distance Learning for Lexical Post-Processing in Handwriting Recognition., IWFHR-9,p.462-467. (2004).
- [4] Milewski, R and Govindaraju, V. Handwriting Analysis of Pre-Hospital Care Reports. IEEE Proceedings. Seventeenth IEEE Symposium on Computer-Based Medical Systems (CBMS) (2004).
- [5] Koerich, A.L.; Leydier, Y.; Sabourin, R.; Suen, C.Y. A Hybrid Large Vocabulary Handwritten Word Recognition System using Neural Networks with Hidden Markov Models. IWFHR-8; 6-8. (2002).
- [6] Xue, H., and Govindaraju, V. On the Dependence of Handwritten Word Recognizers on Lexicons. IEEE Trans. PAMI, Vol. 24, No. 12, p. 1553-1564. (2002).
- [7] Govindaraju, V., Slavik, P., and Xue, H. Use of Lexicon Density in Evaluating Word Recognizers. IEEE Trans PAMI, Vol. 24, No.6, p.789-800. (2002).
- [8] Favata, J: Offline General Handwritten Word Recognition Using an Approximate BEAM Matching Algorithm, IEEE Trans. PAMI, 23 (9): 1009-1021 (2001).
- [9] Zimmermann, M. and Mao, J. Lexicon Reduction using Key Characters in Cursive Handwritten Words. Pattern Recog. Letters; Vol 20, p.1297-1304. (1999).
- [10] Kim, G., and Govindaraju, V.: A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications. IEEE Trans. PAMI 19(4): 366-379 (1997).
- [11] Kaufmann, G.; Bunke, H.; Hadorn, M. Lexicon Reduction in an HMM-Framework Based on Quantized Feature Vectors. Proc. ICDAR 97; Vol. 2 , 18-20, p.1097-1101. (1997).
- [12] Caesar, T.; Gloger, J.M.; Mandler, E. Using Lexical Knowledge for the Recognition of Poorly Written Words. Proc. ICDAR 95; Vol. 2, 14-16, p.915-918. (1995).