Automatic Recognition of Handwritten Medical Forms for Search Engines

Robert Jay Milewski, *Student Member IEEE,* Venu Govindaraju, *Fellow IEEE*

Abstract

A new paradigm, which models the relationships between handwriting and topic categories, in the context of medical forms, is presented. The ultimate goals are (i) the recognition of medical handwriting, and (ii) the use of such information for practical applications such as a medical form search engine. Medical forms have diverse, complex and large lexicons consisting of English, Medical and Pharmacology corpus. Our technique shows that a few recognized characters, returned by handwriting recognition, can be used to construct a linguistic model capable of representing a medical topic category. This allows (i) a reduced lexicon to be constructed, thereby improving handwriting recognition performance, and (ii) PCR forms to be tagged with a topic category and subsequently searched by information retrieval systems. We present an improvement of 8% in raw recognition rate and a search precision of 0.86 at the 0.1 recall position on a data set of unconstrained handwritten medical forms filled in emergency environments.

Index Terms

Handwriting Analysis, Language Models, Pattern Matching, Retrieval Models, Search Process

I. Introduction

This paper describes the first automatic recognition system for handwritten medical forms. In the United States, any pre-hospital emergency medical care provided must be documented. Departments of Health for each state provide a standard medical form to document all patient care from the beginning of the rescue effort until the patient is transported to the hospital. State laws require that emergency personnel fill out one form for each patient.

Figure 1 shows an example Pre-Hospital Care Report (PCR) [48] form which contains 16 information regions (see Table I). Handwriting, from PCR regions 8, 9, 11, 13 and 14 are used for recognition and retrieval analysis. There are two phases to our research: (i) the recognition of handwriting on the medical form, and (ii) a medical form query retrieval engine. Handwriting recognition is used to tag medical forms with a topic category to subsequently

[0] Manuscript received July 17, 2006. This work was supported by the National Science Foundation.

1) Form, agency, and ambulance vehicle Identification	9) Subjective Assessment
2) Patient and physician contact information	10) Presenting Problem
3) Care in progress on arrival and mechanism of injury	11) Past Medical History
4) Dispatch Information	12) Vital/Signs
5) Patient Transfer Information	13) Objective Physical Assessment
6) Rescue times between rescue and transport phases	14) Physical Assessment Extension and/or Comments
7) Extrication and patient vehicle information	15) Treatment Given
8) Chief Complain	16) Ambulance Crew Identification

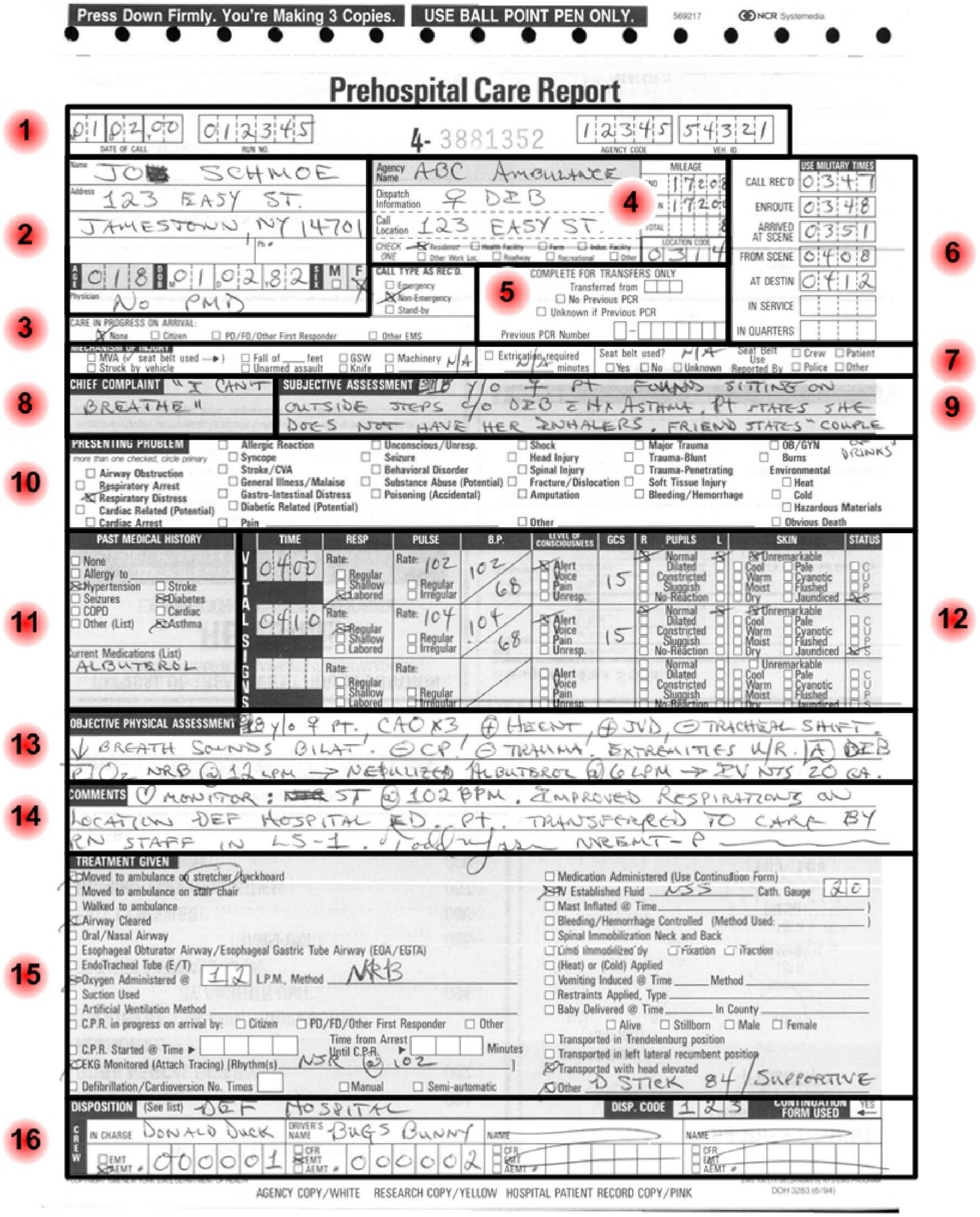
TABLE I
PCR CATEGORIES

improve recognition performance. The medical forms reflect large lexicons containing Medical, Pharmacology and English corpus. While current state of the art recognizers report recognition performance between ∼58-78%, on comparable lexicon sizes in the postal application [27] [49] [50], our experiments show ∼25% raw match recognition performance on the medical forms. This underscores the extremely complicated nature of medical handwriting (Figure 1). We have developed a method of automatically determining the topic category of a PCR form using machine learning and computational linguistics techniques. We demonstrate the strategy for improving the raw word recognition rate by about 8% for a lexicon size of over 5,000 words.

II. BACKGROUND

The basis for reducing the lexicon to improve recognition is a well researched strategy in handwriting recognition [19] [49]. Although handwriting recognition and lexicon pruning/reduction [33] have been researched substantially over the years, many challenges still persist in the offline domain. Word recognition applications range from automated check recognition [26], postal recognition [14], historical documents recognition [15] [18], and now emergency medical documents [35] [36] [37]. Strategic recognition techniques for handwriting algorithms such as Hidden Markov Models (HMM) [28] [34] [38] [22] [9], Artificial Neural Networks (ANN) [40] [4] [11] [16] [10], and Support Vector Machines (SVM) [1] [5] have been developed. Lexicon reduction has been shown to be critical to improvement of performance primarily because of the minimization of possible choices [19]. Even the systems dealing with a large vocabulary corpus have been successful [28] [29].

Additionally, some lexicon reduction strategies have used the extraction of character


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Press Down Firmly. You're Making 3 Copies. USE BALL POINT PEN ONLY. 569217 NCR Systemedia

Prehospital Care Report

DATE OF CALL 01 02 00 RUN NO. 012345 4-3881352 AGENCY CODE 12345 VEH. ID. 54321

Name JO[redacted] SCHMOE
Address 123 EASY ST.
JAMESTOWN, NY 14701
Ph #
AGE 018 DOB 01 02 82 SEX □ M ☒ F
Physician NO PMD

Agency Name ABC AMBULANCE
Dispatch Information ♀ DIB
Call Location 123 EASY ST.
CHECK ONE ☒ Residence □ Health Facility □ Farm □ Indus. Facility □ Other Work Loc. □ Roadway □ Recreational □ Other
MILEAGE END 17208 BEGIN 17200 TOTAL 8
LOCATION CODE 0314

CALL TYPE AS REC'D. □ Emergency ☒ Non-Emergency □ Stand-by

COMPLETE FOR TRANSFERS ONLY
Transferred from
□ No Previous PCR
□ Unknown if Previous PCR
Previous PCR Number

USE MILITARY TIMES

CALL REC'D	0347
ENROUTE	0348
ARRIVED AT SCENE	0351
FROM SCENE	0408
AT DESTIN	0412
IN SERVICE	
IN QUARTERS	

CARE IN PROGRESS ON ARRIVAL: ☒ None □ Citizen □ PD/FD/Other First Responder □ Other EMS

MECHANISM OF INJURY □ MVA (✓ seat belt used →) □ Struck by vehicle □ Fall of ___ feet □ Unarmed assault □ GSW □ Knife □ Machinery □ N/A
□ Extrication required N/A minutes
Seat belt used? N/A □ Yes □ No □ Unknown
Seat Belt Use Reported By □ Crew □ Patient □ Police □ Other

CHIEF COMPLAINT "I CAN'T BREATHE"

SUBJECTIVE ASSESSMENT 18 Y/O ♀ PT FOUND SITTING ON OUTSIDE STEPS C/O DIB c̄ HX ASTHMA. PT STATES SHE DOES NOT HAVE HER INHALERS. FRIEND STATES "COUPLE OF DRINKS"

PRESENTING PROBLEM (more than one checked, circle primary)
□ Airway Obstruction
□ Respiratory Arrest
☒ Respiratory Distress
□ Cardiac Related (Potential)
□ Cardiac Arrest
□ Allergic Reaction
□ Syncope
□ Stroke/CVA
□ General Illness/Malaise
□ Gastro-Intestinal Distress
□ Diabetic Related (Potential)
□ Pain
□ Unconscious/Unresp.
□ Seizure
□ Behavioral Disorder
□ Substance Abuse (Potential)
□ Poisoning (Accidental)
□ Shock
□ Head Injury
□ Spinal Injury
□ Fracture/Dislocation
□ Amputation
□ Other
□ Major Trauma
□ Trauma-Blunt
□ Trauma-Penetrating
□ Soft Tissue Injury
□ Bleeding/Hemorrhage
□ OB/GYN
□ Burns
Environmental
□ Heat
□ Cold
□ Hazardous Materials
□ Obvious Death

PAST MEDICAL HISTORY
□ None
□ Allergy to
☒ Hypertension
□ Seizures
□ COPD
□ Other (List)
□ Stroke
☒ Diabetes
□ Cardiac
☒ Asthma
Current Medications (List) ALBUTEROL

VITAL SIGNS

TIME	RESP	PULSE	B.P.	LEVEL OF CONSCIOUSNESS	GCS	PUPILS R	PUPILS	PUPILS L	SKIN	STATUS
0400	Rate: □ Regular □ Shallow ☒ Labored	Rate: 102 □ Regular □ Irregular	102/68	☒ Alert □ Voice □ Pain □ Unresp.	15	☒ □ □ □ □	Normal Dilated Constricted Sluggish No-Reaction	☒ □ □ □ □	☒ Unremarkable □ Cool □ Pale □ Warm □ Cyanotic □ Moist □ Flushed □ Dry □ Jaundiced	□ C □ U □ P ☒ S
0410	Rate: ☒ Regular □ Shallow □ Labored	Rate: 104 □ Regular □ Irregular	104/68	☒ Alert □ Voice □ Pain □ Unresp.	15	☒ □ □ □ □	Normal Dilated Constricted Sluggish No-Reaction	☒ □ □ □ □	☒ Unremarkable □ Cool □ Pale □ Warm □ Cyanotic □ Moist □ Flushed □ Dry □ Jaundiced	□ C □ U □ P ☒ S
	Rate: □ Regular □ Shallow □ Labored	Rate: □ Regular □ Irregular		□ Alert □ Voice □ Pain □ Unresp.		□ □ □ □ □	Normal Dilated Constricted Sluggish No-Reaction	□ □ □ □ □	□ Unremarkable □ Cool □ Pale □ Warm □ Cyanotic □ Moist □ Flushed □ Dry □ Jaundiced	□ C □ U □ P □ S

OBJECTIVE PHYSICAL ASSESSMENT 18 Y/O ♀ PT., CAO X3, ⊕ HEENT, ⊕ JVD, ⊖ TRACHEAL SHIFT. ↓ BREATH SOUNDS BILAT. ⊖ CP! ⊖ TRAUMA! EXTREMITIES W/R. Ⓐ DIB Ⓟ O₂ NRB @ 12 LPM → NEBULIZED ALBUTEROL @ 6 LPM → IV NTS 20 GA.

COMMENTS ♡ MONITOR: ~~NSR~~ ST @ 102 BPM. IMPROVED RESPIRATIONS ON LOCATION DEF HOSPITAL ED. PT. TRANSFERRED TO CARE BY RN STAFF IN LS-1. Todd NREMT-P

TREATMENT GIVEN
☒ Moved to ambulance on [✓ stretcher]/backboard
□ Moved to ambulance on stair chair
□ Walked to ambulance
☒ Airway Cleared
□ Oral/Nasal Airway
□ Esophageal Obturator Airway/Esophageal Gastric Tube Airway (EOA/EGTA)
□ EndoTracheal Tube (E/T)
☒ Oxygen Administered @ 12 L.P.M., Method NRB
□ Suction Used
□ Artificial Ventilation Method
□ C.P.R. in progress on arrival by: □ Citizen □ PD/FD/Other First Responder □ Other
□ C.P.R. Started @ Time ▶ Time from Arrest Until C.P.R. ▶ Minutes
☒ EKG Monitored (Attach Tracing) [Rhythm(s) NSR @ 102]
□ Defibrillation/Cardioversion No. Times □ Manual □ Semi-automatic
□ Medication Administered (Use Continuation Form)
☒ IV Established Fluid NSS Cath. Gauge 20
□ Mast Inflated @ Time)
□ Bleeding/Hemorrhage Controlled (Method Used:)
□ Spinal Immobilization Neck and Back
□ Limb Immobilized by □ Fixation □ Traction
□ (Heat) or (Cold) Applied
□ Vomiting Induced @ Time ___ Method
□ Restraints Applied, Type
□ Baby Delivered @ Time ___ In County
□ Alive □ Stillborn □ Male □ Female
□ Transported in Trendelenburg position
□ Transported in left lateral recumbent position
☒ Transported with head elevated
☒ Other D STICK 84 / SUPPORTIVE

DISPOSITION (See list) DEF HOSPITAL DISP. CODE 123 CONTINUATION FORM USED YES ←

CREW
IN CHARGE DONALD DUCK □ EMT ☒ AEMT # 000001
DRIVER'S NAME BUGS BUNNY □ CFR ☒ EMT □ AEMT # 000002
NAME □ CFR □ EMT □ AEMT #
NAME □ CFR □ EMT □ AEMT #

AGENCY COPY/WHITE RESEARCH COPY/YELLOW HOSPITAL PATIENT RECORD COPY/PINK DOH 3283 (6/94)

Fig. 1. Pre-Hospital Care Report (PCR) Labeled [48]

information for lexicon reduction, such as that by Guillevic, et al. [20]. However, such strategies reduce the lexicon for a single homogeneous category, namely cities within the country of Finland. In addition, usage of word length estimates for a smaller lexicon are available [20]. Caesar, et al. [6] also state that prior reduction techniques [45] [46] [41] are unsuitable since they can only operate on very small lexicons due to enormous computational burdens [6]. Caesar [6] further indicates that Suen's [47] approach of n-gram combinatorics is sensitive to segmentation issues, a common problem with medical form handwriting [6]. However, Caesar's method [6] and those which are dependent on using the character information, and/or the character information of only one word to directly reduce the lexicon, suffer if one of the characters is selected incorrectly [6]. This is observable in the cursive or mixed-cursive handwriting types.

Many existing schemes, such as that of Zimmermann [51], assume that some characters can be extracted. However, in the medical handwriting domain this task is error prone. Therefore, operating a reduction scheme which can be robust to incorrectly chosen characters is necessary. We use sequences of characters to determine the medical topic category which has a lexicon of its own, thereby reducing the issues of using the character information directly. Similar to the study by Zimmermann et al. [51], the length of words are used with phrases.

Kaufmann, et al. [25] present another HMM strategy which is primarily a distance-based method and uses model assumptions which are not applicable in the medical environment. For example, Kaufmann [25] assumes that "...people generally write more cooperatively at the beginning of the word, while the variability increases in the middle of the word." In the medical environment, variability is apparent when multiple health care professionals enter data on the same form. The medical environment also has exaggerated and/or extremely compressed word lengths due to erratic movement in a vehicle and limited paper space. Kaufmann [25] only provides a reduction of 25% of the lexicon size with little to no improvement in error rate, and the experiments are run only on a small sample of words.

III. Lexicon Reduction

This research proposes the following hypothesis which is verified experimentally: A sequence of confidently recognized characters, extracted from an image of handwritten medical text, can be used to represent a topic category. The construction of medical form training and test decks has been created manually. A software data entry system has been developed which

10 Body Systems: Circulatory/Cardiovascular, Digestive, Endocrine, Excretory, Immune, Integumentary, Musculoskeletal, Nervous, Reproductive, Respiratory.
6 Body Range Locations: Abdomen, Back/Thoracic/Lumbar, Chest, Head, Neck/Cervical, Pelvic/Sacrum/Coccyx.
4 Extremity Locations: Arms/Shoulders/Elbows, Feet/Ankles/Toes, Hands/Wrists/Fingers, Legs/Knees.
4 General: Fluid/Chemical Imbalance, Full Body, Hospital Transfer/Transport, Senses.

TABLE II

Categories are denoted by these Anatomical Positions

allows human truthers to segment all PCR form regions and words, and provide a human interpretation for the word, denoted as the truth. Truthing is done in two phases: (i) the digital transcription of medical form text, and (ii) the classification of forms into topic categories. The distribution of PCR forms under each category is approximately equal in both the training and test decks (see Table II). The task has been supervised and performed by a health care professional with several years of field emergency medical services (EMS) experience. This emergency medical data set is the first of its kind.

A PCR can be tagged with multiple categories. In our data set, no form had more than five category tags. The subjectivity involved in determining the categories makes the construction of a hierarchical chart representing all patient scenarios with respective prioritized anatomical regions a difficult task and exceeds the scope of this research. The following are some examples for classifying medical form text into categories (see Table II):

Example 1: A patient treated for an emergency related to her pregnancy would be included in the *Reproductive System* category (see Table II).

Example 2: A conscious and breathing patient treated for gun shot wounds to the abdominal region would fall into the *Circulatory/Cardiovascular System* due to potential loss of blood, as well as being categorized for *Abdominal, Back, and Pelvic* categories (see Table II).

We take characters with the highest recognition as an input and produce higher level topic categories. A knowledge base is constructed during the *training phase* from a set of PCR forms. The knowledge base contains the relationships between terms and categories. The *recognition phase* takes an unknown form, and reduces the lexicon using the knowledge base. This phase is evaluated using a separate testing deck. Finally, after all content of the PCR form has been recognized, a search can take place by entering in a query. This phase is tested by querying the system with a deck of phrase inputs. The forms are then ranked accordingly and returned to the user.

In the training phase, a mechanism for relating uni-grams and bi-grams (henceforth: uni/bi-grams) as well as categories from a PCR training deck are constructed. The testing phase then evaluates the algorithm's ability to determine the categories from a test form by using a lexicon driven word recognizer (LDWR) [27] to extract the top-choice uni/bi-gram characters from all words. A maximum of two characters per word are considered, given that LDWR [27] successfully extracts a bi-gram with spatial encoding information 40% of the time. If ≥ 3 characters are selected, then LDWR [27] successfully extracts a character $\leq 1\%$ of the time. Hence the maximum value of n in the n-grams is taken to be 2 (see examples in Figure 4).

A. Training

The training stage involves a series of steps to construct a matrix that represents relationships between terms and categories. Each form can have up to five categories. In the first phase, lexicons are constructed using all the words from all forms under a category. In the second phase, phrases are extracted from the form using a cohesion equation. These phrases are then converted to ESI encoding terms (ESI denotes "Exact Spatial Information" used as the encoding procedure for the uni/bi-gram terms; see definitions later in this section). A matrix is then constructed utilizing the ESI terms for the rows and the categories in the columns. The matrix is then normalized, weighted, and prepared in Singular Value Decomposition format.

A list of about 400 stopwords provided by PubMed are omitted from the lexicon [39] [21]. An additional list of about 50 words (e.g. male, female, etc.) found in most PCR's, which have little bearing on the category are omitted from the cohesion analysis (the frequency of two words co-occurring versus occurring independently; see Equation 1) but retained in the final lexicon. It is common to apply other filters to reduce the likelihood of morphological mismatches [21]. However, strategies such as 'stemming' [21] cannot be applied before recognition because the

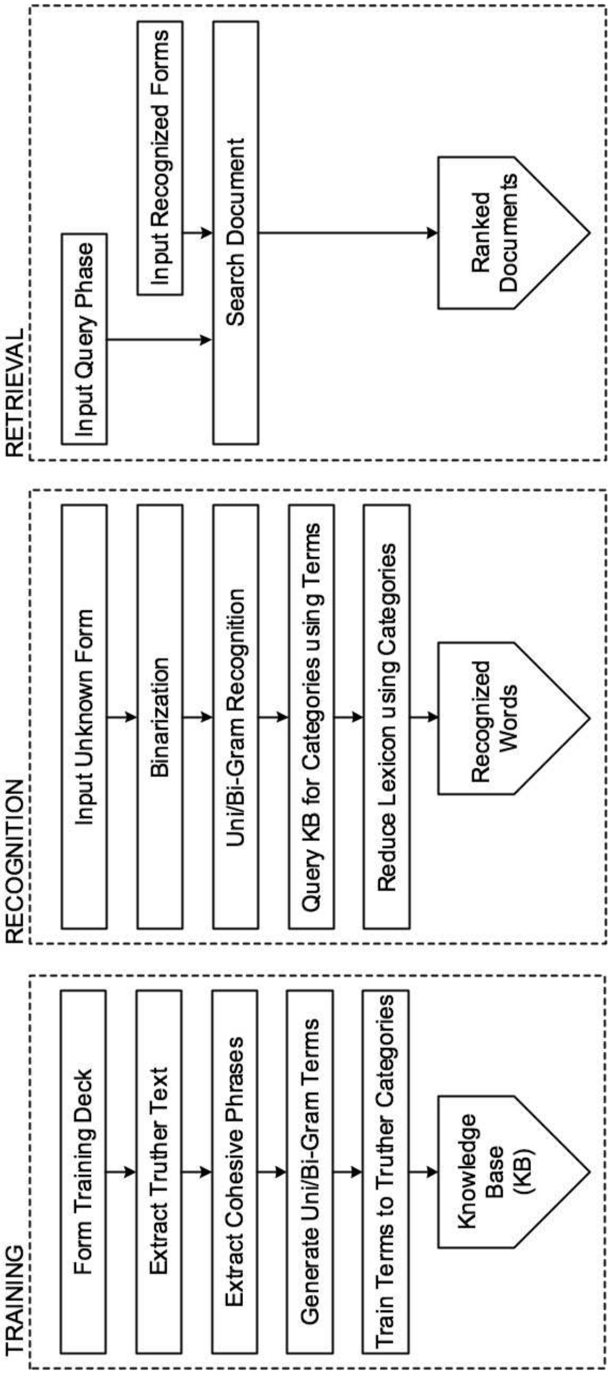


Fig. 2. Proposed Algorithm Road Map

text is not yet ASCII and is therefore unknown. Consider a handwritten word image representing "rhythms" that needs to be recognized. The alteration of "rhythms" to "rhythm" in the lexicon will affect recognition performance. However, at the end of classification, these words are considered equivalent. Therefore, word stemming is applied after the LDWR [27] has determined the ASCII word translation.

A passage P is the set of all words w for a PCR form under a category C treated as a single string. For each C, every pair of passages, denoted P_1 and P_2, is compared. A phrase is defined as a sequence of adjacent non-stopwords [13]. Here we denote w_x as a word located at position *x* within a passage P. If $w_a \in P_1, w_a^{'} \in P_2, w_b \in P_1, w_b^{'} \in P_2$ such that $b^{'} > a^{'}$ and $b > a$, then a potential phrase consisting of exactly two words is constructed. The cohesion of phrases under each C is then computed. If the cohesion is above a threshold, then that phrase represents that category C. Thus a category C is represented by a sequence of high cohesion phrases using only those PCR passages manually categorized under C.

$$cohesion(w_a, w_b) = z \bullet \frac{f(w_a, w_b)}{\sqrt{f(w_a)f(w_b)}} \tag{1}$$

The cohesion between any two words w_a and w_b is computed by the frequency that w_a and w_b occur together versus existing independently. The top 40 cohesive phrases are retained for each category (see Equation 1).

Consider the following two unfiltered text sentences S_1 and S_2 under the category *legs*:

S_1: "right femur fracture"

S_2: "broken right tibia and femur"

The candidate phrases CP_1 and CP_2 after the filtering step are:

CP_1: "right femur" ... "right fracture" ... "femur fracture"

CP_2: "broken right" ... "right femur" ...

The phrase "right femur" is computed from CP_1 and CP_2, given that w_a and $w_a^{'}$ = "right", w_b and $w_b^{'}$ = "femur", and the conditions $b > a$ and $b' > a'$ have been met. If the cohesion for "right femur" is above the threshold across all PCR forms under the *legs* category, then this phrase is retained as a representative of the category *legs*.

FREQUENCY	COHESION	PHRASE
6	0.67	DCAP BTLS
166	0.35	CHEST PAIN
91	0.38	PAIN 0
1860	2.49	PAIN HIP
144	0.34	HIP JVD
112	0.39	PAIN CHANGE
275	0.81	HIP FX
110	0.37	HIP CHANGE
82	0.38	PAIN 10
163	0.40	JVD PAIN
106	0.40	CAOX3 PAIN
202	0.50	PAIN JVD
213	0.55	PAIN LEG
205	0.42	CHEST HIP
3	0.33	PERPENDICULAR DECREASE
121	0.33	FELL HIP
118	0.36	PAIN FX
2251	3.01	HIP PAIN
390	0.83	PAIN CHEST
288	0.59	HIP CHEST

TABLE III

TOP COHESIVE PHRASES FOR THE CATEGORY: PELVIS

Tables III and IV illustrate some top choice cohesive phrases generated. Digestive system and pelvic region are anatomically *close*. However, different information is reported in these two cases resulting in mostly different cohesive phrases. Those which are the same, such as *CHEST PAIN* have different cohesion values. This implies that it is likely that the term frequencies will also be different and therefore commonly occurring terms need to be weighted appropriately to their categories (this will be discussed in more detail later). Phrases sometimes may not make sense by themselves, however, this is the result of using a cohesive phrase formula in which words may not be adjacent.

There are three strategies for term representations: NSI, ESI and ASI. These terms will later be modeled to an anatomical category and used as the essential criterion for lexicon reduction.

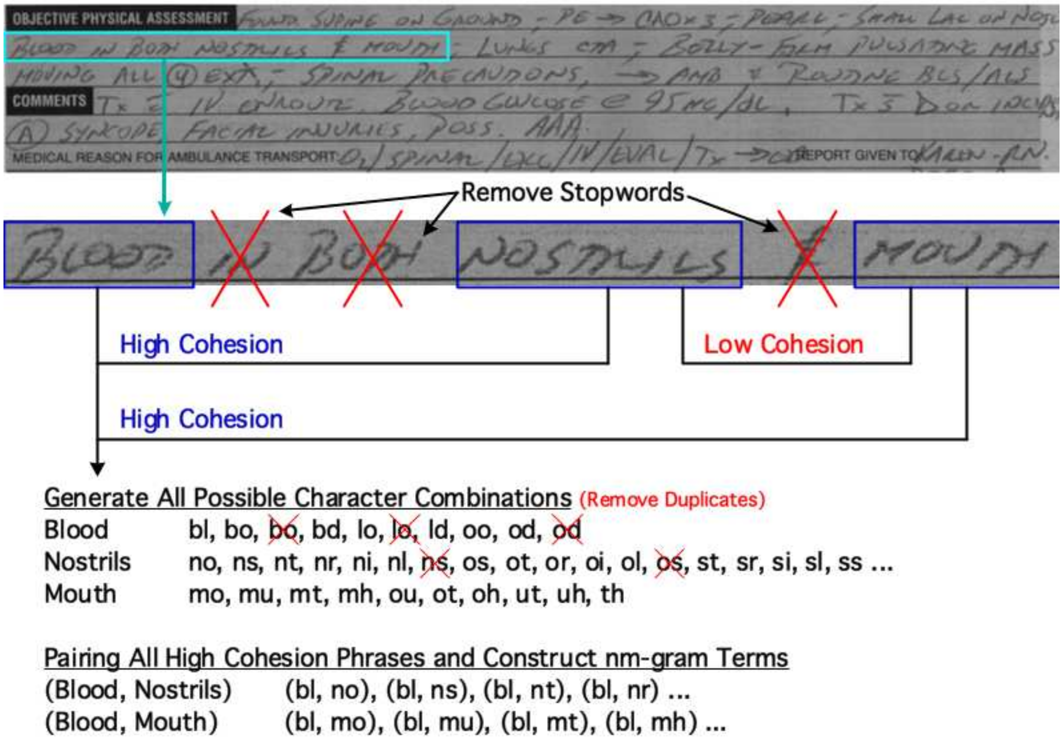

OBJECTIVE PHYSICAL ASSESSMENT
COMMENTS
MEDICAL REASON FOR AMBULANCE TRANSPORT
REPORT GIVEN TO
Remove Stopwords
High Cohesion
Low Cohesion
High Cohesion
Generate All Possible Character Combinations (Remove Duplicates)
Blood bl, bo, ~~bo~~, bd, lo, ~~lo~~, ld, oo, od, ~~od~~
Nostrils no, ns, nt, nr, ni, nl, ~~ns~~, os, ot, or, oi, ol, ~~os~~, st, sr, si, sl, ss ...
Mouth mo, mu, mt, mh, ou, ot, oh, ut, uh, th
Pairing All High Cohesion Phrases and Construct nm-gram Terms
(Blood, Nostrils) (bl, no), (bl, ns), (bl, nt), (bl, nr) ...
(Blood, Mouth) (bl, mo), (bl, mu), (bl, mt), (bl, mh) ...

Fig. 3. Term Extraction from High Cohesive Phrases

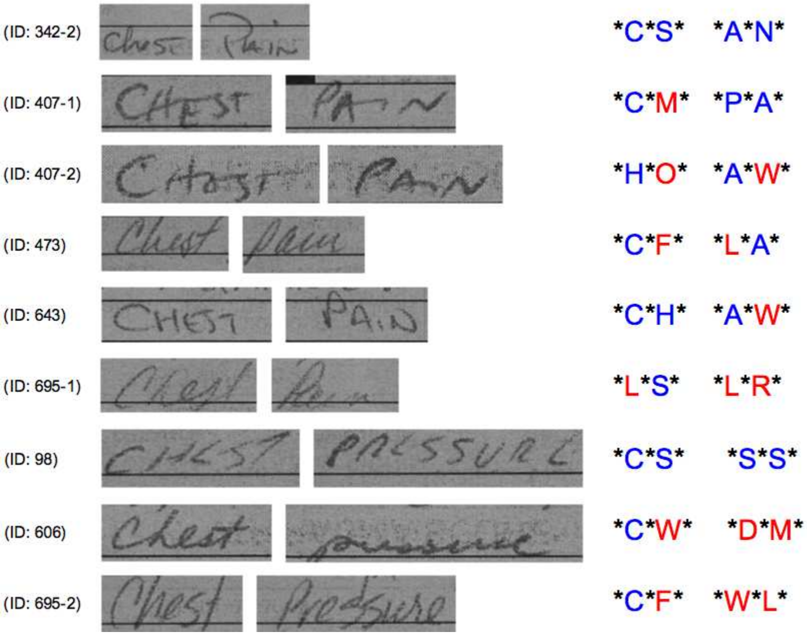


Fig. 4. NSI Encodings Example (Blue Letters: LDWR[27] successfully extracted)

FREQUENCY	**COHESION**	**PHRASE**
30	0.72	PAIN INCIDENT
5	0.31	PAIN TRANSPORTED
42	0.54	PAIN CHEST
52	0.81	STOMACH PAIN
9	0.25	HOME PAIN
6	0.43	VOMITING ILLNESS
39	0.51	CHEST PAIN
4	0.24	CHEST SOFT
25	0.54	PAIN SBM
31	0.37	PAIN X4
31	0.47	PAIN JVD
11	0.34	PAIN EDEMA
25	0.44	PAIN PMSX4
6	0.21	PAIN SOFT
3	0.21	SBM INCIDENT
11	0.25	PAIN LEFT

TABLE IV
TOP COHESIVE PHRASES FOR THE CATEGORY: *Digestive System*

No Spatial Information (NSI):

An asterisk (*) indicates that zero or more characters are found between C_1 and C_2. NSI encodings are the most simple form of encoding (see Figure 4 examples).

UNI-GRAM ENCODING: $*C*$

BI-GRAM ENCODING: $*C_1 * C_2*$

BI-GRAM ENCODING EXAMPLE: BLOOD → *L*D*

Exact Spatial Information (ESI):

The integers (x, y, z) represent the precise number of characters between C_1 and C_2. ESI encodings are an extension of the NSI encodings with the inclusion of precise spatial information. In other words, the number of characters before, after and between the highest confidence C_1 and C_2 characters are part of the encoding. These encodings are the most successful in our experiments since there are fewer term collisions involved. Hence the ESI encodings are preferred.

UNI-GRAM ENCODING: xCy

BI-GRAM ENCODING: xC_1yC_2z

BI-GRAM ENCODING EXAMPLE: BLOOD → 1L2D0

Approximate Spatial Information (ASI):

The integers (x_a, y_a, z_a), denoted as length codes, represent an estimated range of characters between C_1 and C_2. A '0' indicates no characters, a '1' indicates between one and two characters, and a '2' represents greater than 2 characters. The ASI encodings are an approximation of ESI encodings designed to handle cases when the precise number of characters is not known with high confidence.

UNI-GRAM ENCODING: x_aCy_a

BI-GRAM ENCODING: $x_aC_1y_aC_2z_a$

BI-GRAM ENCODING EXAMPLE: BLOOD → 1L1D0

Combinatorial Analysis

The quantity of all possible NSI, ESI and ASI uni-gram and bi-gram combinations, for a given word of character length n, such that $n \geq 1$, is represented by Equation 2. Regardless of the encoding, the same quantity of combinations exists since the distance between characters is known.

$$\mathcal{C}(n) = \left(\left(\sum_{i=1}^{n-1} (n-i) \right) + n \right) = \left(\left(\left(\frac{n}{2} \right) (n-1) \right) + n \right) \tag{2}$$

However, the function $\mathcal{C}$ only considers the combinations of an individual entry. The combination inflation of a uni/bi-gram phrase is shown by Equation 3. The equation parameters *a* and *b* represent the string lengths of the words considered in a phrase.

$$\mathcal{P}(a,b) = \mathcal{C}(a) \cdot \mathcal{C}(b) \tag{3}$$

For example:

Let the phrase to evaluate uni/bi-gram combinations be *PULMONARY DISEASE*.

Let n = length(“PULMONARY”) = 9

Let m = length(“DISEASE”) = 7

$\mathcal{C}$(n) = 45 uni-gram + bi-gram combinations for “PULMONARY”

$\mathcal{C}$(m) = 28 uni-gram + bi-gram combinations for “DISEASE”

$\mathcal{P}$(n,m) = 1,260 uni-gram + bi-gram phrase combinations for *PULMONARY DISEASE*

Each of these encodings has its advantages and disadvantages. The choice is ultimately based on the quality of the handwriting recognizer’s (LDWR) ability to extract characters. If the handwriting recognizer cannot successfully extract positional information, then NSI is the best approach. If extraction of positional information is reliable, then the ESI is the best approach. However, NSI and ASI create more possibilities for confusion since distances are either approximated or omitted. ESI is more restrictive on the possibilities as the precise spacing is used leading to lesser confusion among terms.

Using the ESI protocol, all possible uni/bi-gram terms are synthetically extracted from each cohesive phrase under each category. For example, BLOOD can be encoded to the uni-gram 0B4 (zero characters before ’B’ and four characters after ’B’) and the bi-gram 0B3D0 (zero characters before ’B’, three characters between ’B’ and ’D’ and zero characters following ’D’). All possible synthetic positional encodings are generated for each phrase and chained together (a ’$’ is used to denote a chained phrase). For example, CHEST PAIN encodes to: 0C4$0P0A2 ... 0C4$1A2 ... 0C0H3$0P1I1 ... 0C0H3$0P2N0, etc. Therefore, each category now has a list of encoded phrases consisting of positional encoded uni/bi-grams. These terms are the most primitive representative links to the category used throughout the training process. In the training phase, the synthetic information can be extracted since the text is known. However, in the testing phase, a recognizer will be used to automatically produce an ESI encoding since the test text is not known. To improve readability, the notation (W_1, W_2) is used to represent an ESI encoding of a two-word phrase (e.g. Myocardial Infarction: (my, in), (my, if), (my, ia), etc ...).

A matrix A, of size $|T|$ by $|C|$, is constructed such that the rows of the matrix represent the set of terms T, and the columns of the matrix represent the set of category C. The value at matrix coordinate (t,c) is the frequency that each term is associated with the category. The term frequency corresponds to the phrasal frequency from which it was

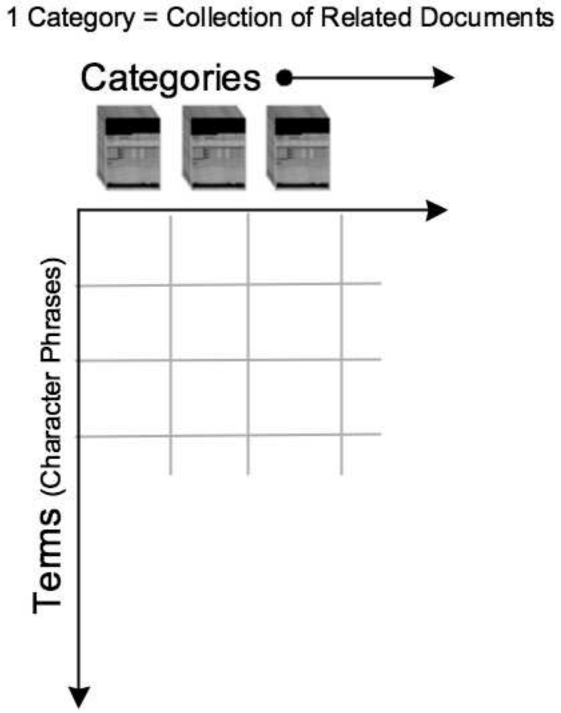

1 Category = Collection of Related Documents
Categories
Terms (Character Phrases)

Fig. 5. Term Category Matrix (TCM) Overview

derived. It is the same value as the numerator in the cohesion formula (refer to Equation 1): $f(w_a, w_b)$. For example, if the frequency of CHEST PAIN is 50, then all term encodings generated from CHEST PAIN, such as (ch, pa), will also receive a frequency of 50 in the matrix.

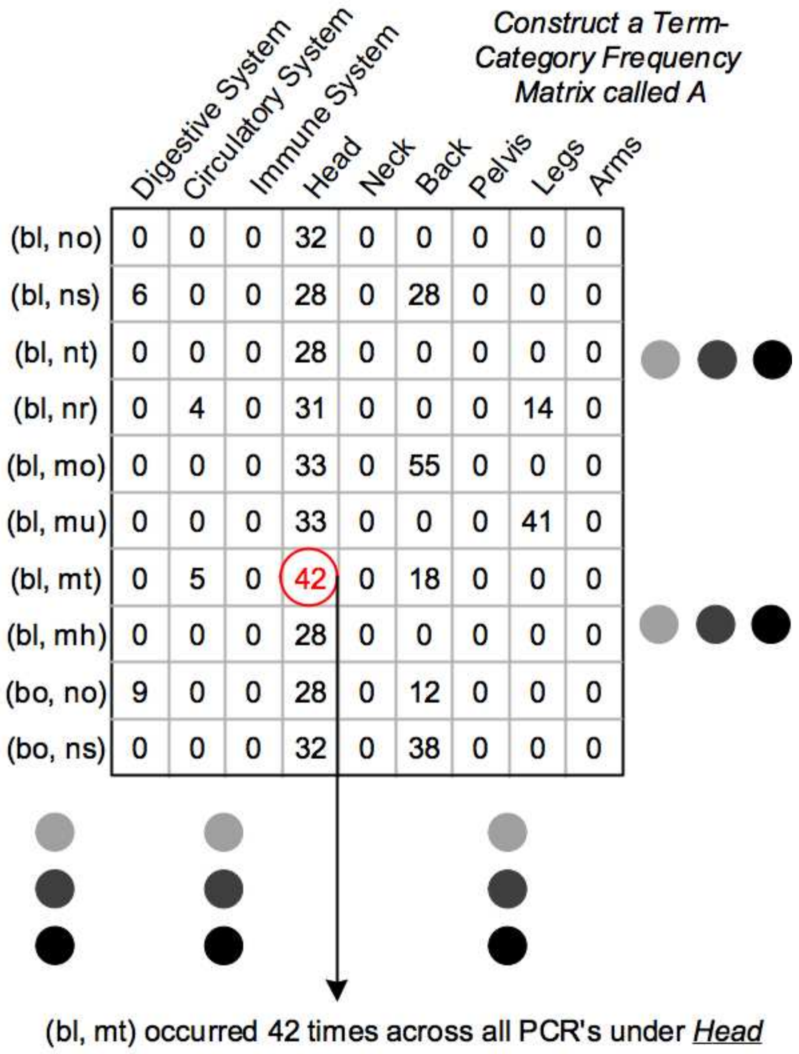
Step 1: Compute the normalized matrix B from A using Equation 4 [7] [8]:

$$B_{t,c} = \frac{A_{t,c}}{\sqrt{\sum_{e=1}^{n} A_{t,e}^2}} \tag{4}$$

Matrix A is the input matrix containing raw frequencies, Matrix B is the output matrix with normalized frequencies, and (t,c) is a (term, category) coordinate within a matrix.

Step 2: Term Discrimination Ability

The Term Frequency times Inverse Document Frequency (TF x IDF) is used to favor those terms which occur frequently with a small number of categories as opposed to their existence in all categories [31] [44]. While Luhn [31] asserts that medium frequency terms would best



Construct a Term-Category Frequency Matrix called A
(bl, mt) occurred 42 times across all PCR's under *Head*

	Digestive System	Circulatory System	Immune System	Head	Neck	Back	Pelvis	Legs	Arms
(bl, no)	0	0	0	32	0	0	0	0	0
(bl, ns)	6	0	0	28	0	28	0	0	0
(bl, nt)	0	0	0	28	0	0	0	0	0
(bl, nr)	0	4	0	31	0	0	0	14	0
(bl, mo)	0	0	0	33	0	55	0	0	0
(bl, mu)	0	0	0	33	0	0	0	41	0
(bl, mt)	0	5	0	42	0	18	0	0	0
(bl, mh)	0	0	0	28	0	0	0	0	0
(bo, no)	9	0	0	28	0	12	0	0	0
(bo, ns)	0	0	0	32	0	38	0	0	0

Fig. 6. TCM Frequency Construction Example

resolve a document, it precludes classification of rare medical words. Salton's [44] theory, stating that terms with the most discriminatory power are associated with fewer documents, allows a rare-medium frequent word to resolve the document.

STEP 2A Compute the weighted matrix X from B using Equation 5 [7] [8] [21]:

$$IDF(t) = log_2 \frac{n}{c(t)} \tag{5}$$

IDF gives the inverse-document-frequency on term *t*, where c(t) is the number of categories containing term *t*.

Step 2B Weight the normalized matrix by IDF values using Equation 6 [7] [8] [23] [21]:

$$X_{t,c} = IDF(t) \cdot B_{t,c} \tag{6}$$

Matrix B is the normalized matrix from Step 1, IDF is the computational step defined in Step 2, and Matrix X is a normalized and weighted matrix.

The normalized and weighted term-category matrix can now be used as the knowledge base for subsequent classification. A singular value decomposition variant, which incorporates a dimensionality reduction step allows a large term-category matrix to represent the PCR training set (see Equation 7). This facilitates a category query from an unknown PCR using the LDWR [27] determined terms [7] [8] [12].

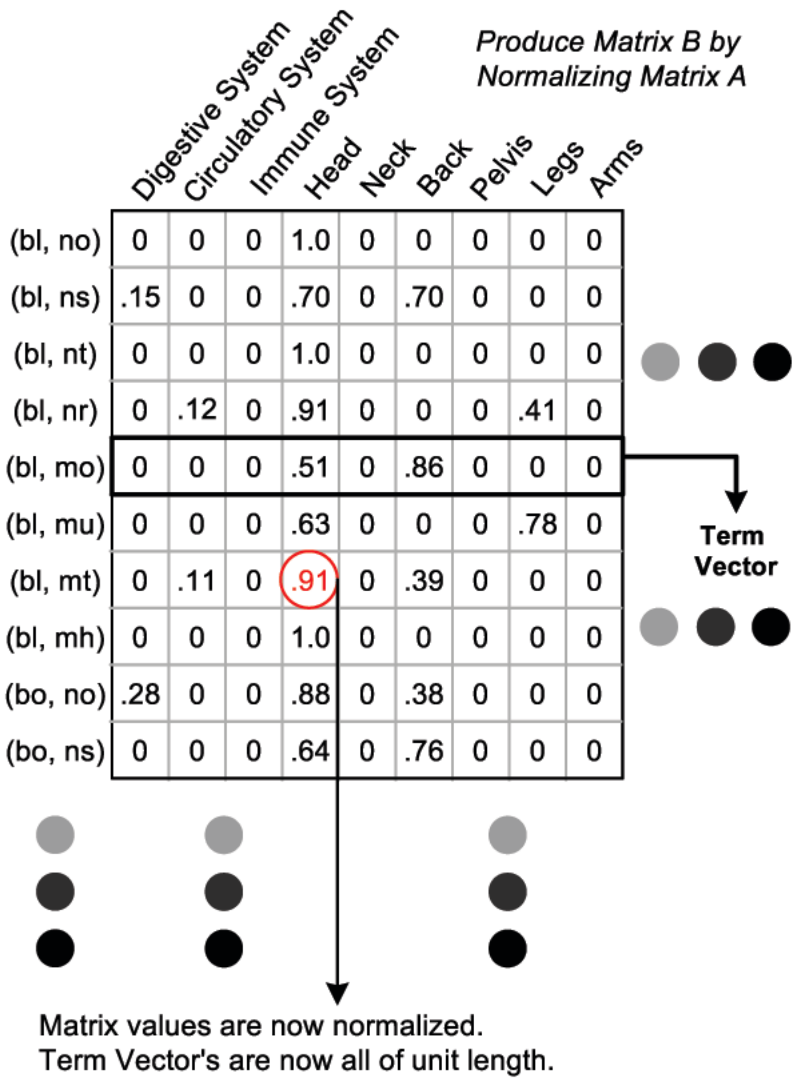
$$X = U \bullet S \bullet V^T \tag{7}$$

Matrix X is decomposed into 3 matrices: U is a (T x k) matrix representing term vectors, S is a (k x k) matrix, and V is a (k x C) matrix representing the category vectors.

The value k represents the number of dimensions to be finally retained. If k equals the targeted number of categories to model, then SVD is performed without the reduction step. Therefore, in order to reduce the dimensionality, the condition $k < |C|$ is necessary to reduce noise [12].

B. Testing

Given an unknown PCR form, the task is to determine the category of the form, and use the reduced lexicon associated with the determined category to drive the word recognizer, LDWR [27]. In addition, the category determined can be used to tag the form which can be subsequently used for information retrieval. The query task is divided into the following steps:



Produce Matrix B by Normalizing Matrix A

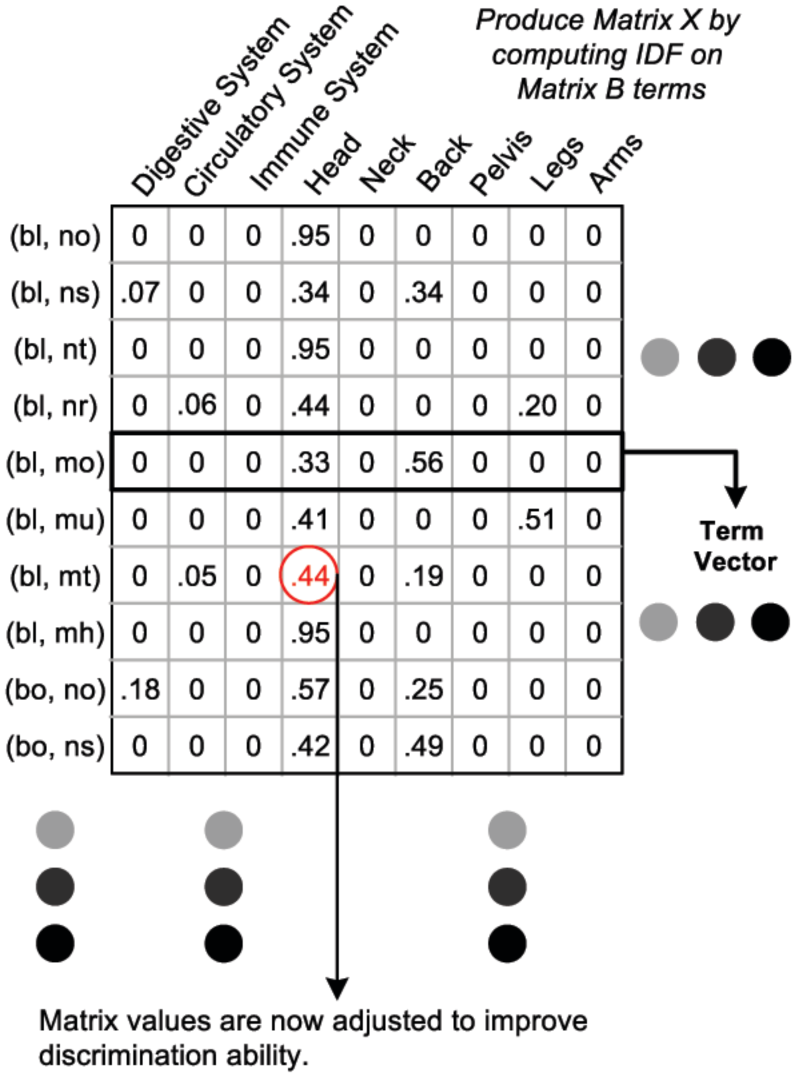
	Digestive System	Circulatory System	Immune System	Head	Neck	Back	Pelvis	Legs	Arms
(bl, no)	0	0	0	1.0	0	0	0	0	0
(bl, ns)	.15	0	0	.70	0	.70	0	0	0
(bl, nt)	0	0	0	1.0	0	0	0	0	0
(bl, nr)	0	.12	0	.91	0	0	0	.41	0
(bl, mo)	0	0	0	.51	0	.86	0	0	0
(bl, mu)	0	0	0	.63	0	0	0	.78	0
(bl, mt)	0	.11	0	.91	0	.39	0	0	0
(bl, mh)	0	0	0	1.0	0	0	0	0	0
(bo, no)	.28	0	0	.88	0	.38	0	0	0
(bo, ns)	0	0	0	.64	0	.76	0	0	0

Matrix values are now normalized.
Term Vector's are now all of unit length.

Fig. 7. TCP Normalization

(i) Term Extraction, (ii) Pseudo-Category Generation, and (iii) Candidate Category Selection [7] [8].

Given a new PCR image, all image words are extracted from the form, and LDWR [27] is used to produce a list of confidently recognized characters for each word. These are used to encode the positional uni/bi-grams consistent with the format during training. All combinations of uni/bi-phrases in the PCR form are constructed. Each word has exactly one uni-gram and exactly one bi-gram. A phrase consists of exactly two unknown words. Therefore it is



Produce Matrix X by computing IDF on Matrix B terms

	Digestive System	Circulatory System	Immune System	Head	Neck	Back	Pelvis	Legs	Arms
(bl, no)	0	0	0	.95	0	0	0	0	0
(bl, ns)	.07	0	0	.34	0	.34	0	0	0
(bl, nt)	0	0	0	.95	0	0	0	0	0
(bl, nr)	0	.06	0	.44	0	0	0	.20	0
(bl, mo)	0	0	0	.33	0	.56	0	0	0
(bl, mu)	0	0	0	.41	0	0	0	.51	0
(bl, mt)	0	.05	0	.44	0	.19	0	0	0
(bl, mh)	0	0	0	.95	0	0	0	0	0
(bo, no)	.18	0	0	.57	0	.25	0	0	0
(bo, ns)	0	0	0	.42	0	.49	0	0	0

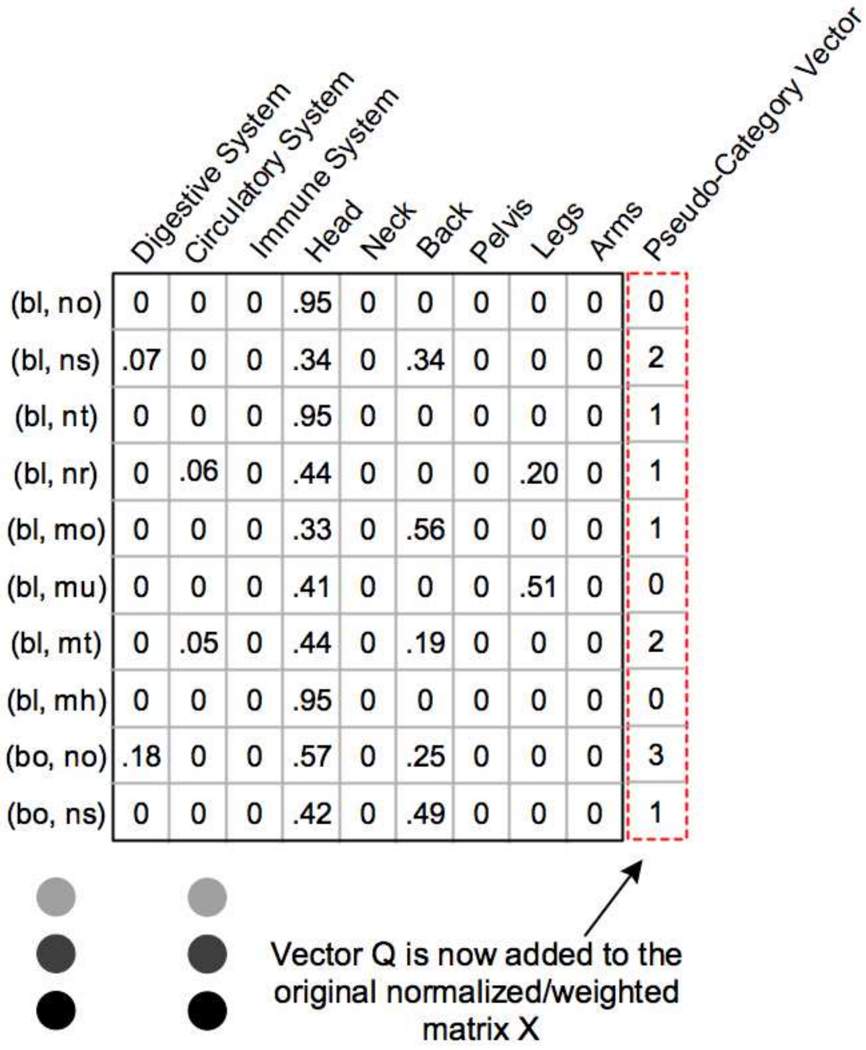
Term Vector

Matrix values are now adjusted to improve discrimination ability.

Fig. 8. TCM Inverse Document Frequency (IDF)

represented by precisely four uni/bi-phrases (BI-BI, BI-UNI, UNI-BI and UNI-UNI).

A *(m x 1)* query vector Q is derived, which is then populated with the term frequencies for the generated sequences from the Term-Extraction step. If a term is not encountered in the training set, then it is not considered. Positional bi-grams are generated to yield the trained terms 37% of the time, and similarly positional uni-grams 57% of the time. The experiments here illustrate this to be a sufficient number of terms. A scaled vector representation of Q is



	Digestive System	Circulatory System	Immune System	Head	Neck	Back	Pelvis	Legs	Arms	Pseudo-Category Vector
(bl, no)	0	0	0	.95	0	0	0	0	0	0
(bl, ns)	.07	0	0	.34	0	.34	0	0	0	2
(bl, nt)	0	0	0	.95	0	0	0	0	0	1
(bl, nr)	0	.06	0	.44	0	0	0	.20	0	1
(bl, mo)	0	0	0	.33	0	.56	0	0	0	1
(bl, mu)	0	0	0	.41	0	0	0	.51	0	0
(bl, mt)	0	.05	0	.44	0	.19	0	0	0	2
(bl, mh)	0	0	0	.95	0	0	0	0	0	0
(bo, no)	.18	0	0	.57	0	.25	0	0	0	3
(bo, ns)	0	0	0	.42	0	.49	0	0	0	1

Vector Q is now added to the original normalized/weighted matrix X

Fig. 9. Pseudo-Category Vector

then produced by multiplying Q^T and U.

Once the pseudo document is derived, R-SVD is applied for the following reasons: (i) It converts the query into a vector space compatible input, and (ii) the dimensional reduction can help reduce noise [12]. Since the relationship between terms and categories is scaled by variance, the reduction allows parametric removal of less significant term-category relationships.

The task is now to compare the pseudo-category vector Q with each vector in $V_r \bullet S_r$ (from

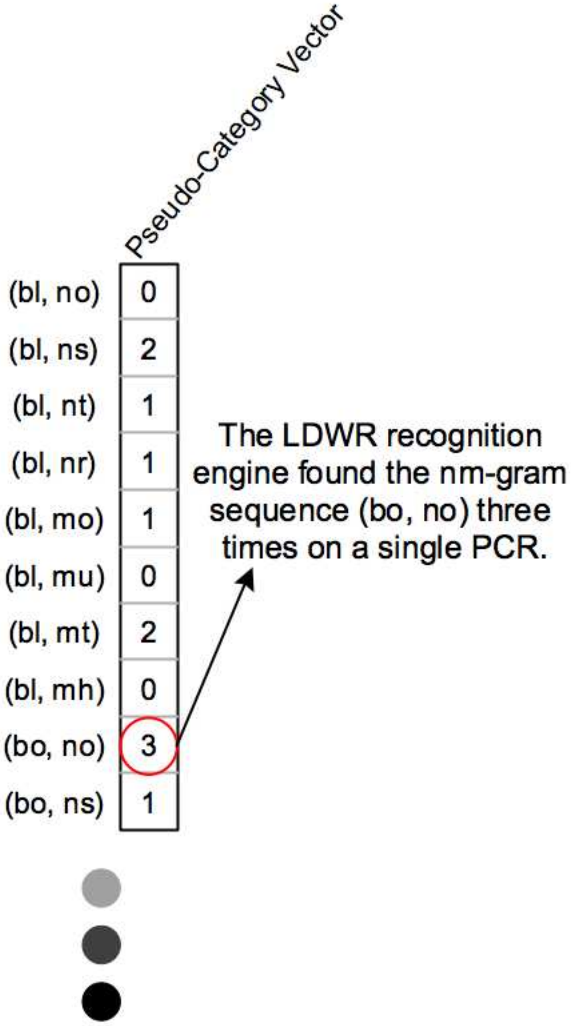


Fig. 10. Pseudo-Category Integration

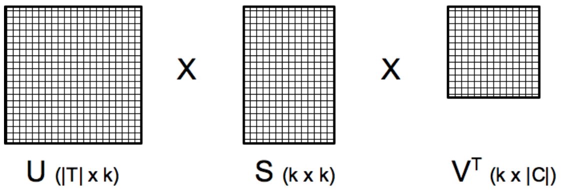


Fig. 11. Matrix Decomposition Visual

the training phase) using a scoring mechanism. The cosine rule is used for matching the query [7] [8]. Both *x* and *y* are dimensional vectors used to compute the cosine in Equation 8. Vectors *x* and *y* in the equations represent the comparison of the vectors: pseudo-document Q with every column vector in $V_r \bullet S_r$.

$$z = cos(x, y) = \frac{x \cdot y^T}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} \tag{8}$$

Each cosine score is mapped onto a sigmoid function using the least square fitting method, thereby producing a more accurate confidence score [7] [8]. The least squares regression line used to satisfy the equation f(x) = ax + b are shown in Equations 9 and 10 [30]:

$$a = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \tag{9}$$

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i) \tag{10}$$

The fitted sigmoid confidence is produced using the cosine score and the regression line, using equation (9):

$$confidence(a, b, z) = \frac{1}{1 + e^{-(az+b)}} \tag{11}$$

The confidence scores are then used to rank the categories. If a category is above an empirically chosen threshold, then that category is retained for the PCR. Multiple categories may be thus retained. All words corresponding to the selected categories are then used to construct a new lexicon which is finally submitted to the LDWR recognizer [27]. Given a test PCR form, and the reduced lexicon, the LDWR [27] converts the handwritten medical words to ASCII.

Each word which is recognized is compared with the truth. However, a simple string comparison is insufficient due to spelling mistakes and root variations of word forms which are semantically identical. This occurs 20% of the time within the test deck words. Therefore, Porter stemming [42] [24] [43] and Levenshtein String Edit Distance [2] of 1 allowable penalty are performed on both the truth and the recognizer result before they are compared. Levenshtein

FIGHT vs EIGHT vs LIGHT	FINE vs FIRE
MEDICAL vs MEDICATION	FOOD vs FOOT
1400 vs 2400	LEFT vs LIFT
BAIL vs RAIL	MOANING vs MORNING
BALL vs CALL	MARK vs MARY
MOLE vs MOVE	PUNCH vs LUNCH
CALF vs CALL	REACH vs REACT
CARD vs CARE vs CART	SCARE vs CARE
COLD vs TOLD	SEVER vs FEVER
NECK vs DECK	STABLE vs TABLE
FALL vs CALL	FEET vs FEED
FOUND vs BOUND vs SOUND vs POUND	

TABLE V

WORD COLLISIONS

	CL	**CLT**	**AL**	**ALT**	**SL**	**SLT**	**RL**	**RLT**
ACC	76.34%	76.92%	63.52%	66.59%	70.51%	71.51%	70.70%	71.06%
ERR	71.93%	69.65%	57.24%	47.12%	62.26%	59.44%	62.04%	59.45%
RAW	23.31%	25.32%	32.31%	41.73%	30.30%	32.73%	30.62%	32.63%
TLS	5,628	8,156	1,193	1,246	2,514	2,620	2,401	2,463
!L	-	-	23.89%	8.02%	16.06%	10.46%	16.61%	12.23%
!HL	-	-	33.33%	97.98%	48.19%	73.99%	46.59%	62.96%

TABLE VI

HANDWRITING RECOGNITION PERFORMANCE

is only applied to a word that is believed to be ≥ 4 characters in length. For example, PAIN and PAINS are identical. However, this also results in an improper comparison in about 11% of the corrections (see Table V).

IV. EXPERIMENTS

Our training data consists of 750 PCR forms and the test data consists of a separate blind set of 62 PCR forms. In all experiments it is assumed that the word segmentation and extraction has been performed by a person. Also, forms in which 50% of the content is indecipherable by

	CLT to RLT	**CL to RL**	**CLT to ALT**	**CLT to SLT**
RAW Match Rate	↑ 7.48%	↑ 7.42%	↑ 17.58%	↑ 7.42%
Error Rate	↓ 10.78%	↓ 10.88%	↓ 24.53%	↓ 10.21%

TABLE VII

COMPARISON BETWEEN HANDWRITING RECOGNITION EXPERIMENTS

LEXICON ANALYSIS METRIC	**VALUE**
Accuracy of Reduction (α)	0.33
Degree of Reduction (ρ)	0.83
Reduction Efficacy (η)	0.06
Lexicon Density (ϱ')	1.07 → 0.87
Lexicon Density (ϱ")	0.50 → 0.78

TABLE VIII

LEXICON REDUCTION PERFORMANCE BETWEEN THE COMPLETE LEXICON (CL) AND THE REDUCED LEXICON (RL)

ENVIRONMENT ITEM	**VALUE**
Training Deck PCR Size	750
Testing Deck PCR Size	62
Training Deck Lexicon Size	5,628
Testing Deck Lexicon Size	2,528
Training + Testing Deck Lexicon Size	8,156
Training Deck Words for Modeling	42,226
Testing Deck Words to Recognize	3,089
Modeled Categories / RSVD Dimensions	24
Category Selection Threshold	0.55
Maximum Categories per Form	5
Average Categories per form	1.40
Max Phrases Per Category	50
Apple OS X Memory Usage	520 MB
Apple OS X G4 1GHZ Train Time	15-20 mins/exp
Apple OS X G4 1GHZ Test Time	3 hrs/exp

TABLE IX

HANDWRITING RECOGNITION SYSTEM ENVIRONMENT

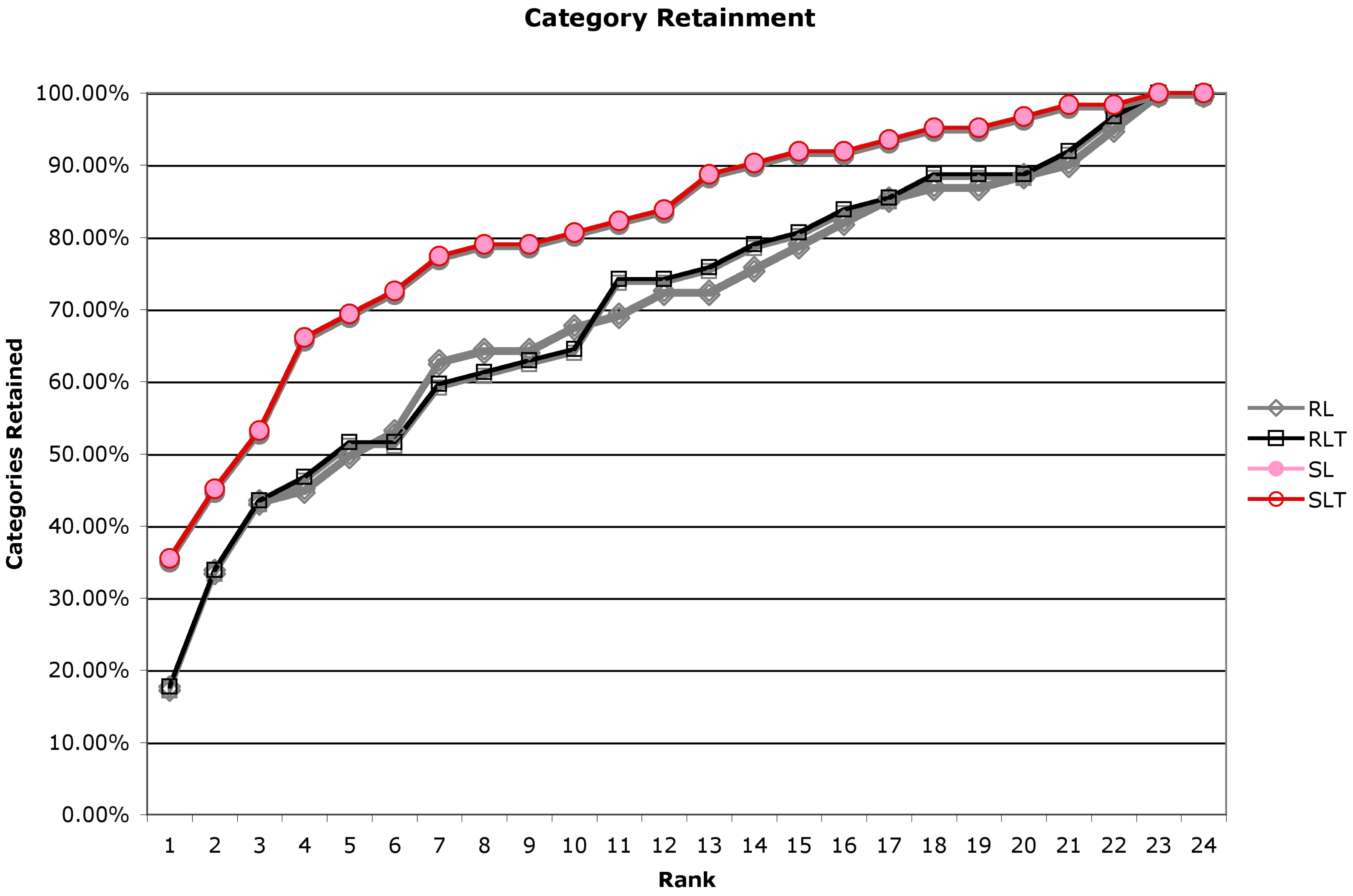


Fig. 12. Category Retainment by Rank

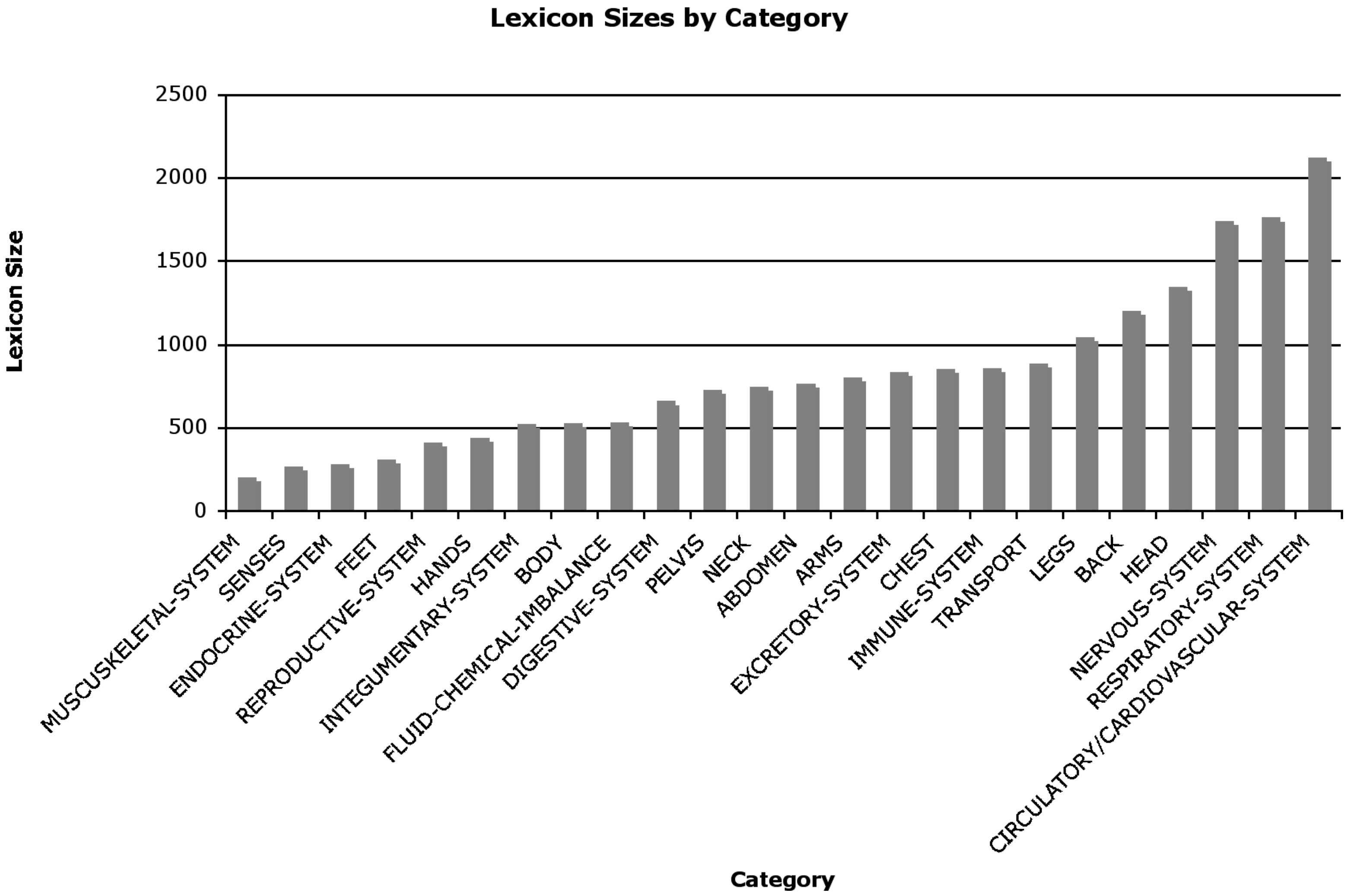


Fig. 13. Sorted Lexicon Sizes by Category

a human being are omitted. This occurs 15% of the time.

CL (complete training lexicon): The union of all words in the training set.
CLT (complete training lexicon + test deck lexicon): The union of all the words in the training and test sets.

AL (assumed training lexicon): This is a reduced lexicon constructed from the training deck where the categories are determined by an Oracle.
ALT (assumed training lexicon + test deck lexicon): Same as AL except that all words from the test set are also inserted into the training deck category lexicon. This gives the upper bound for the effectiveness of the reduced lexicon strategy.

RL (reduced lexicon): The reduced lexicon from the training deck, which is the union of words from the top ranked categories returned by the word recognizer. This is a practical measure of the current performance of the system.
RLT (reduced lexicon + test deck lexicon): Same as RL except that all words from the test set are inserted into the training deck category lexicon. This shows the effectiveness of word recognition under the assumption that the category lexicons are complete.

SL (synthetic term generation): This is the reduced lexicon in which the categories are determined by a synthetic generation of the truth word. This is the theoretical upper bound of RL in which the handwriting recognition is a 100% accept rate with a 0% error rate.
SLT (synthetic term generation + test deck lexicon): Same as SL except that all words from the test set are inserted into the training deck category lexicon. This is the theoretical upper bound of RLT.

ACC (accept recognition rate): number of words the word recognizer accepts above an empirically decided threshold.
ERR (error recognition rate): number of words incorrectly recognized among the accepted words.
RAW (raw recognition rate): top choice word recognition rate without use of thresholds.
!L (truther word not present in the lexicon): percentage of words (for a specific experiment) not

in the lexicon as a result of incorrectly chosen categories or due to the absence of that word in the training deck.

!HL (human being could not completely decipher word): percentage of the !L set in which even human beings could not reliably decipher all or some of the characters in the word (given the context).

In reference to Table VII which is computed from the most relevant changes in Table VI : The theoretical RLT (i.e. comparing RLT to CLT) improves the RAW match rate by 7.48% and drops the error rate 10.78% with a *degree of reduction* ρ = 61.59%. The practical RL (i.e. comparing RL to CL) improves the RAW match rate by 7.42% and drops the error rate by 10.88%. The RLT and RL numbers are close due to the difference in the initial lexicon sizes: CLT/RLT starts with 6,561 words (i.e. training deck and testing deck lexicons) whereas the CL/RL starts with 5,029 words (i.e. training deck lexicon only). The RLT lexicon is more complete, but the lexicon is larger. The RL lexicon is less complete, but the lexicon is smaller. Thus, RLT gives the advantage that the recognizer has a greater chance of the word being a possible selection and RL gives the advantage of the lexicon being smaller. The ALT shows the theoretical upper bound for the paradigm: (i) the categories are correctly determined 100%, and (ii) the lexicon is complete. The ALT (i.e. comparing ALT to CLT) improves the RAW match rate by 17.58% and drops the error rate 24.53% with a *degree of reduction* ρ = 83.01%. The synthetic experiments (SL and SLT) also do not offer much improvement which shows perfect character extraction does not guarantee recognition improvement. This is due to two reasons: (i) a form is a representation of many characters and so some incorrectly recognized characters are tolerated, and (ii) the remaining words on the form to be recognized are difficult to determine even when the lexicon is constructed with only the words of known uni/bi-gram terms.

Table VIII provides insight into the effectiveness of the lexicon reduction from the complete lexicon (CL) to the reduced lexicon (RL) experiments. The performance measures for lexicon reduction as described by Madhvanath [32] and Govindaraju, et al. [19] are used with alteration to the definition of reduction efficacy. The *Accuracy of Reduction* $\alpha = E(\mathcal{A})$, such that $\alpha \in [0, 1]$ [32], and $\mathcal{A}$ is a random variable [3], indicates the existence of the truth in the lexicon. The function *E* computes the expectation [3]. The *Degree of Reduction* $\rho = E(\mathcal{R})$, such that $\rho \in [0, 1]$ [32], represents the mean size of the reduced lexicon. The *Reduction Efficacy* $\eta = \Delta_{LDWR} \times \alpha^{1-\rho}$, such that Δ_{LDWR}, $\eta, \alpha, \rho \in [0, 1]$, is a measure of the effectiveness of a lexicon with respect to a lexicon driven recognizer. This formula is defined differently in this research to weigh the importance of accuracy over the reduction and include the reductions effect on the recognizer. The larger the efficacy value is, the better is the effectiveness of the reduction for one recognizer versus another. The larger the *Lexicon Density* $\varrho_{LDWR}(\mathcal{L}) = (\upsilon_{LDWR}(\mathcal{L}))(f_{LDWR}(n) + \delta_{LDWR})$ value (such that $\upsilon_{LDWR}(\mathcal{L}) = \frac{n(n-1)}{\sum_{i \neq j} d_{LDWR}(\omega_i, \omega_j)}$ and $d_{LDWR}(\omega_i, \omega_j)$ is a recognizer dependent computation used to denote a distance metric between two supplied words) the more *similar* or *close* the lexicon words are [19]. A supplemental distance measure denoted by the *N-Gram Lexicon Distance Metric* $d_{LDWR}(\omega_i, \omega_j) = \gamma(\omega_i, \omega_j)/\Gamma(\omega_i, \omega_j)$, introduced in this research and substituted into the lexicon density equation ϱ, provides a measure of uni/bi-grams existing within the lexicon. The value γ represents the number of uni/bi-gram terms that are *not* common between ω_i and ω_j. Γ denotes the total number of uni/bi-gram term combinations between ω_i and ω_j. In order to distinguish between the *lexicon density distance metric* and the *n-gram lexicon distance metric* equations, the values ϱ' and ϱ'' will be respectively used. The *lexicon density distance metric* $\varrho^{'}$ shows less confusion among lexicon words considering all the characters are equally important. This implies that the reduced lexicon will be less confusing to the recognizer. The *n-gram lexicon distance metric* shows an increase in the quantity of words with common NSI encodings. This implies the recognizer has a greater chance of selecting a word using the confidently selected characters.

V. SEARCH METHOD

In this section, various search engine approaches are compared. The inputs to the search engine are a set of PCR medical forms and a query. The output is the set of forms which match the input query.

Search engines are mostly based on the assumption that the text is already in a digital text format. The technologies have focused on parsing and organizing the content in a variety of formats (e.g. PDF, PS, HTML, XML, and other proprietary document formats). There is no widely used search engine technology which can directly search and analyze the content of digital handwritten documents. This query ability is important in Health Surveillance applications to access medical forms by simply offering a query.

In order to have a query deck of sufficient size, we use the leave-1-out strategy which is explained as follows. Suppose a total of 10 PCR's are available. Take the first PCR as the test deck and the remaining 9 PCR's as the training deck, and perform the recognition and tagging on that single PCR. Next, repeat the process, except that now the test deck consists only of the 2nd PCR while the training consists of the first PCR and the remaining 8 PCR's. The recognition and tagging on the 2nd PCR is now performed. This exhaustive processing of recognition and tagging repeats 10 times, thereby providing a training deck and an unbiased test deck of the same size. Applying this process to 800 PCR forms, the notion is the same, except the split is Leave-100; i.e. 8 experiments are performed using groups of 100. Finally, a set of 1,250 phrases, constructed from adjacent non-stopwords, are extracted from a blind deck of 200 PCR forms (i.e. these 200 forms are not a subset of the 800 deck) such that each phrase is found in at least one form in the 800 deck.

A query is performed by scanning the forms in the 800 test deck for recognized words that match an input query phrase. Two query experiments are performed and displayed in Figure 14: CL and RL. In the CL (complete lexicon) experiment, the raw LDWR recognized words computed from the full lexicon are compared against the query. In the RL (reduced lexicon) experiment, the raw LDWR words computed from the reduced lexicon are compared against the query. A set of ranking rules are applied, relevance determined, and the recall-precision table generated (see Table VI and Figure 14). A relevant PCR is a document in which a human truther classifies at least one occurrence of each word from the input phrase.

Ranking rules given an input phrase of exactly two words are as follows:

- Both words must match the recognized words or that PCR is not returned.
- A double precision rank is computed by summing the values in these two steps:
  - ○ Summing the frequencies of the occurring phrase words in the document.
  - ○ Summing the distance between all recognized word occurrences in the document using

Equation 12. Let $d(a_i, b_j)$ be a function which computes the distance between the input phrase of two words, a_i and b_j such that *i* and *j* respectively represent the word position in the document.

$$d(a_i, b_j) = \frac{1}{|a_i - b_j|} \tag{12}$$

Unlike typical text retrieval systems, the words on a PCR may be incorrectly recognized by the handwritten recognition engine. In addition, general search engines need to be concerned about external influences such as spamming, which is not a concern in this application.

Therefore, a more trivial ranking measure such as of nearness/proximity in Equation 12 is sufficient.

The comparison of the complete and reduced lexicon queries can be found in Figure 14. The plot illustrates only those queries which returned at least one record. While the CL retrieval appears to be effective, queries in that series returned 0 forms 73% of the time, and returned only 1 record on average. The RL returned 0 forms 23% of the time and returned 7.5 documents on average. That is, 50% more of the queries generated at least one response. Therefore, the lexicon reduction strategy offers considerable improvement over the complete lexicon strategy. This also illustrates that a small improvement in handwriting recognition rates can offer a huge improvement in search performance.

An alternative search engine approach involving the expansion of the query terms into their respective ESI combinations can be applied directly to the initial LDWR character recognition results. This would effectively bypass the more elaborate search engine except that this alternative approach significantly under-performs. While results are returned 99.8% of the time, with 125 records returned on average, the precision of the results is very low. As intuitively expected, the uni/bi-grams match more terms due to the loss of word information. The recall/precision chart in Figure 15 illustrates a drop in retrieval effectiveness. This demonstrates the dependence of the searches to operation at the word level, rather than at the character level. The lexicon reduction strategy which improves the handwriting recognition performance also improves the search effectiveness as expected.

For example, consider input query phrase *CHEST PAIN*:
CHEST is decomposed into: CH, CE, CS, CT, HE, HS, HT, ES, ET, C, H, E, S, and T.
PAIN is decomposed into: PA, PI, PN, AI, AN, IN, P, A, I, and N.

In addition, the spatial information is known since the input query is provided by a user.

The ESI encodings for *CHEST* is decomposed into: 0C0H3, 0C1E2, 0C2S1, 0C3T0, 1HE2, 1H1S1, 1H2T0, 2E0S1, 2E1T0, 0C4, 1H3, 2E2, 3S1, and 4T0.

The ESI encoding for *PAIN* is decomposed into: 0P0A2, 0P1I1, 0P2N0, 1A0I1, 1A1N0, 2I0N0,

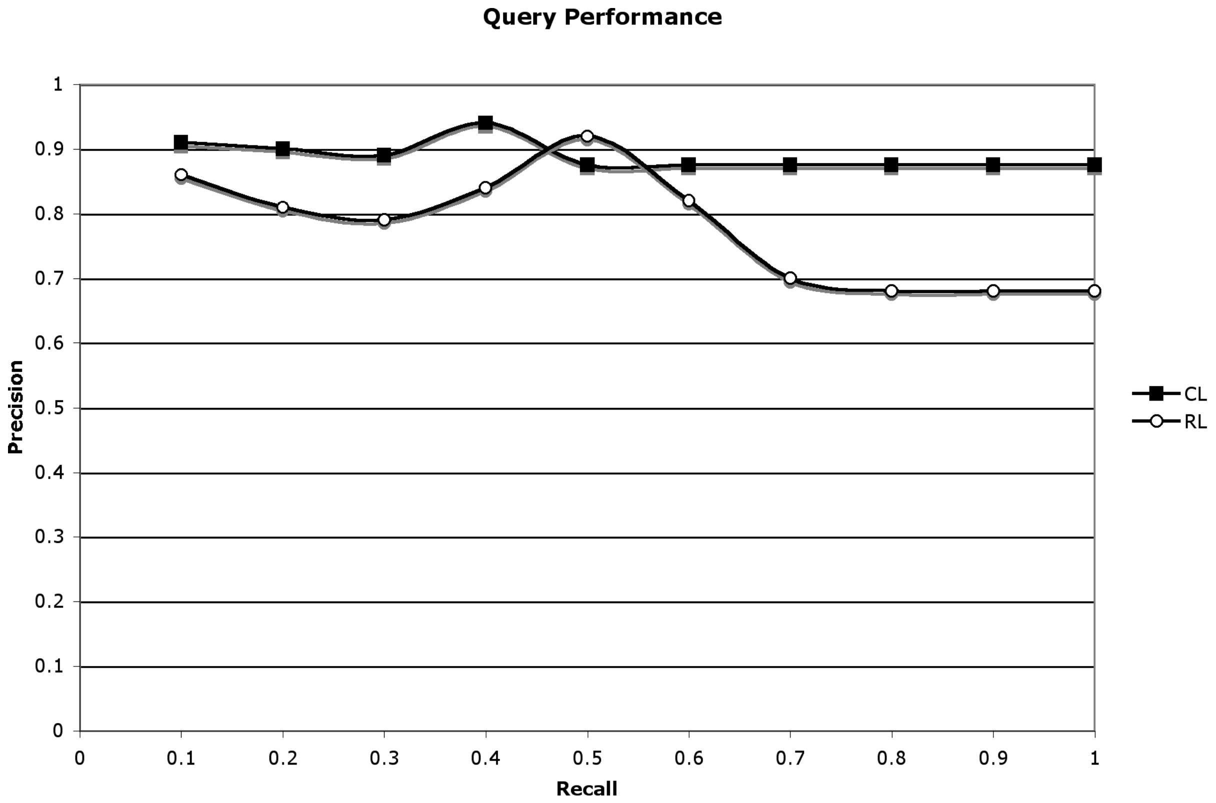


Fig. 14. Recall/Precision Chart for Medical Form Search Engine

0P3, 1A2, 2I1, and 3N0.

Finally, all possible ESI sequences are generated: 0C0H3$0P0A2, 0C0H3$0P1I1, 0C0H3$0P2N0, 0C0H3$1A0I1, etc...

If any of these ESI sequences match any of the character spatial encodings from the LDWR recognition, then that form is returned. Relevancy is determined if the input query words *CHEST* and *PAIN* are actually found on that form according to the truth.

VI. Conclusions

This paper defines a new paradigm for lexicon reduction and information retrieval in the complex situation of handwriting recognition of medical forms. An improvement in raw recognition rate from about 25% of the words on a PCR form to approximately about 33% has

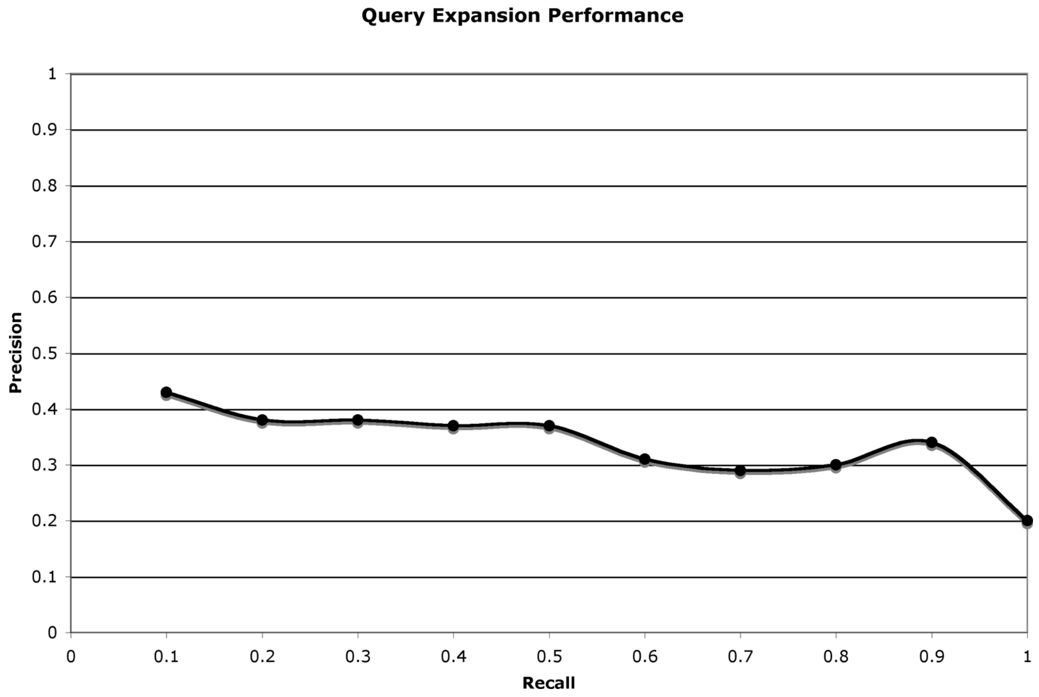


Fig. 15. Recall/Precision Chart using Query Expansion

been shown with a reduction in false accepts by about 7%, a reduction in error rate by about 10%-25%, and a lexicon reduction from 32%-85%. The addition of a category driven query facilitates almost 86% relevant searches at the first recall position in a search engine experiment with medical forms.

Interestingly, certain computational elements of bootstrapping, described in our work, are consistent with the human interpretation of ambiguous handwriting using contextual cues. Our methodology accomplishes this by modeling character terms as a higher level semantic concept which mimics the human ability to recognize a word within context, when some characters are unknown.

ACKNOWLEDGMENT

Special acknowledgements to (i) the National Science Foundation (NSF) for providing funding for this project, to (ii) Casey Czamara, of the Western Regional Emergency Medical Services (WREMS) program operating under the New York State Department of Health for providing necessary resources [48].

REFERENCES

[1] Bahlmann, C., Haasdonk, B., Burkhardt, H. On-Line Handwriting Recognition with Support Vector Machines - A Kernel Approach. International Workshop on Frontiers in Handwriting Recognition. 2002.

[2] Black, P.E., ed. ”Levenshtein Distance”. Algorithms and Theory of Computation Handbook; CRC Press LLC, from Dictionary of Algorithms and Data Structures, NIST, 1999.

[3] Blum, J.R., Rosenblatt, J.I. Probability and Statistics. Chapter 4: Random Variables and Their Distributions; Chapter 6: Expectations, Moment Generating Functions, and Quantiles. W.B. Saunders Company. USA, 1972.

[4] Blumenstein, M., Verma, S. A Neural Based Segmentation and Recognition Technique for Handwritten Words. IEEE International Conference on Neural Networks. 1998.

[5] Byun, H., Lee, S.W. Applications of Support Vector Machines for Pattern Recognition: A Survey. Lecture Notes in Computer Science; Springer. 2002.

[6] Caesar, T., Gloger, J.M., Mandler, E. Using Lexical Knowledge for the Recognition of Poorly Written Words. Third International Conference on Document Analysis and Recognition. Volume 2. pp. 915-918. 1995.

[7] Chu-Carroll, J., and Carpenter, B. Dialogue Management in Vector-Based Call Routing. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. pp. 256-262, 1999.

[8] Chu-Carroll, J., and Carpenter, B., Vector-Based Natural Language Call Routing. Computational Linguistics. Vol. 25, No. 3, pp. 361–388, 1999.

[9] Chen, M.Y., Jundu, A., Zhou, J. Off-Line Handwritten Word Recognition Using a Hidden Markov Model Type Stochastic Network. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1994.

[10] Cho, S.B., Kim, J.H. Applications of Neural Networks to Character Recognition. Pattern Recognition. 1991.

[11] Cho, S.B. Neural-Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals. IEEE Transactions on Neural Networks. 1997.

[12] Deerwester, S., Dumais, S.T., Furnas, G.Q., Landauer, and, T.K., Harshman, R. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. 41(6):391-407, 1990.

[13] Fagan, J. The Effectiveness of a Non-Syntactic Approach to Automatic Phrase Indexing for Document Retrieval. Journal of the American Society for Information Science, 40: 115-132. 1989.

[14] Favata, J.T. Offline General Handwritten Word Recognition Using an Approximate BEAM Matching Algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI); 23 (9); pp1009-1021. 2001.

[15] Feng, S.L. and Manmatha, R. Classification Models for Historic Manuscript Recognition. Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR) 2005.

[16] Gader, P.D., Keller, J.M., Krishnapuram, R., Chiang, J.H. Neural and Fuzzy Methods in Handwriting Recognition. 1997.

[17] Golub, G.B., Van Loan, C.E. Matrix Computations. 2nd Edition. John Hopkins University Press, 1989.

[18] Govindaraju, V., Xue, H. Fast Handwriting Recognition for Indexing Historical Documents. First International Workshop on Document Image Analysis for Libraries (DIAL). 2004.

[19] 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.

[20] Guillevic, D., Nishiwaki, D., and Yamada, K. Word Lexicon Reduction by Character Spotting. Seventh International Workshop on Frontiers in Handwriting Recognition. Amsterdam. 2000.

[21] Hersh, W.R. Information Retrieval: A Health and Biomedical Perspective. 2nd Edition. Springer-Verlag, New York, Inc. USA. 2003.

[22] Hu, J. Brown, M.K., Turin, W. HMM Based Online Handwriting Recognition. IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI). 1996.

[23] Jones, K.S. A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation, 28(1):11-20 1972.

[24] Jones, K.S. and Willet, P. Readings in Information Retrieval, San Francisco: Morgan Kaufmann. 1997.

[25] Kaufmann, G.; Bunke, H.; Hadorn, M. Lexicon Reduction in an HMM-Framework Based on Quantized Feature Vectors. Proc. International Conference on Document Analysis and Recognition (ICDAR) 97; Vol. 2 , 18-20, p.1097-1101. 1997.

[26] Kim, G., Govindaraju, V. Bank Check Recognition using Cross Validation between Legal and Courtesy Amounts. International Journal on Pattern Recognition and Artificial Intelligence. 1996.

[27] Kim, G., and Govindaraju, V.: A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications. IEEE Trans. PAMI 19(4): 366-379. C1997.

[28] Koerich, A.L., Sabourin, R., Suen, C.Y. Fast Two-Level HMM Decoding for Large Vocabulary Handwriting Recognition. International Workshop on Frontiers in Handwriting Recognition. 2004.

[29] Koerich, A.L., Sabourin, R., Suen, C.Y. Large Vocabulary Off-Line Handwriting Recognition: A Survey. Pattern Analysis and Applications. 2003.

[30] Larson, Hostetler, Edwards. Calculus with Analytic Geometry. Chapter 13 : Section 13.9. Fifth Edition. D.C. Heath and Company. USA. 1994.

[31] Luhn, H. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1: 309-317. 1957.

[32] Madhvanath, S. The Holistic Paradigm in Handwritten Word Recognition and its Application to Large and Dynamic Lexicon Scenarios. University at Buffalo Computer Science and Engineering. Ph.D. Dissertation, 1997.

[33] Madvanath, S., Krpasundar, V., Govindaraju, V. Syntactic Methodology of Pruning Large Lexicons in Cursive Script Recognition. Journal. The Journal of the Pattern Recognition Society. Pattern Recognition 34 (2001); Elsevier Science. 2001.

[34] Marti, U.V., Bunke, H. Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition Systems. World Scientific Series in Machine Perception and Artificial Intelligence Series. 2001.

[35] Milewski, R. and Govindaraju, V. Medical Word Recognition using a Computational Semantic Lexicon. Eighth International Workshop on Frontiers in Handwriting Recognition. Canada. 2002.

[36] Milewski, R and Govindaraju, V. Handwriting Analysis of Pre-Hospital Care Reports. IEEE Proceedings. Seventeenth IEEE Symposium on Computer-Based Medical Systems (CBMS). 2004.

[37] Milewski, R and Govindaraju, V. Extraction of Handwritten Text from Carbon Copy Medical Forms. Document Analysis Systems (DAS). Springer-Verlag. 2006.

[38] Nakai, M., Akira, N., Shimodaira, H., Sagayama, S. Substroke Approach to HMM-Based On-Line Kanji Handwriting Recognition. Sixth International Conference on Document Analysis and Recognition. 2001.

[39] National Library of Medicine. PubMed Stop List.

[40] Oh, I.S., Suen, C.Y. Distance Features for Neural Network-Based Recognition of Handwritten Characters. International Journal on Document Analysis and Recognition. 1998.

[41] Okuda, T. Tanaka, E. Kasai, T. A Method for the Correction of Garbled Words based on the Levenshtein Distance. IEEE Transactions on Computers, Col. C-25, No. 2. 1976.

[42] Porter, M.F. An Algorithm for Suffix Stripping. Program, 14: 130-137. 1980.

[43] Rijsbergen, C.J. van, Robertson, S.E. and Porter, M.F. New Models in Probabilistic Information Retrieval. London: British Library. 1980.

[44] Salton, G. Introduction to Modern Information Retrieval. New York. McGraw-Hill. 1983.

[45] Sinha, R.M.K., Prasada, B. Visual Text Recognition through Contextual Processing. Pattern Recognition Vol 21, No 5. pp.463-479. 1988

[46] Srihari, S.N., Hull, J.J., Choudhari, R. Integrating Diverse Knowledge Sources in Text Recognition. ACM Transactions on Office Information Systems. Vol.1 No.1 pp. 68-87. 1983.

[47] Suen. C.Y. N-gram Statistics for Natural Language Understanding and Processing. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 1, No.2, pp 164-172. 1979.

[48] Western Regional Emergency Medical Services. Bureau of Emergency Medical Services. New York State (NYS) Department of Health (DoH). Prehospital Care Report v4.

[49] Xue, H., and Govindaraju, V. Stochastic Models Combining Discrete Symbols and Continuous Attributes - Application in Handwriting Recognition. Proceedings of 5th IAPR International Workshop on Document Analysis Systems. pp. 70-81. 2002.

[50] 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.

[51] Zimmermann, M. and Mao, J. Lexicon Reduction using Key Characters in Cursive Handwritten Words. Pattern Recog. Letters; Vol 20, p.1297-1304. 1999.