Diabetes Differential Diagnosis Application System- a Case Study

Shweta Sheel, MBBS[1], Veronica Heredia, MD, Aman Kumar

Abstract— **There is a significant increase in the number of Type-2 diabetic patients in the past decade. This is mainly due to rapidly rising numbers of obese and overweight people. According to the latest CDC estimates over 7 million people in the US are still undiagnosed of this disease. In this study we developed a Diabetes Differential Diagnosis Application (DDDxA) system that takes textual data from a potential patient as input and based on Natural Language Processing techniques suggests if a person should be recommended for further screening for diabetes or not. The DDDxA system has the potential to perform an initial diagnosis of the patient and provide initial treatment in the field away from a doctor's clinic.**

I. INTRODUCTION

According to a CDC (Center for Disease Control and Prevention) 2011 study (http://www.cdc.gov/diabetes/pubs/estimates11.htm#1) there are 25.8 million people, or 8.3% of the U.S. population, that have diabetes. Out of these 25.8 diabetic patients 18.8 million people are diagnosed with this disease while over 7.0 million people in the US are still undiagnosed. These statistics and estimates are derived from various data systems of the Centers for Disease Control and Prevention (CDC), the Indian Health Service's (IHS), National Patient Information Reporting System (NPIRS), the U.S. Renal Data System of the National Institutes of Health (NIH), the U.S. Census Bureau, and published studies.

There is a need in this field for a semi-automated computerized system that quickly and accurately diagnoses a diabetic condition without the help of specialized or expert physicians. This can allow further referrals and rapid treatment that will allow patients to recover faster than is possible without specialists present.

In this study we have developed a preliminary Diabetes Differential Diagnosis Application (DDDxA) that can identify early diabetic (Type-2) instances almost as accurately as a clinician. The fundamental goal of the DDDxA is to allow patient (by themselves or with a healthcare professional) to enter Patient Symptoms on a Mobile Device connected via a Secure Network to a back end Diagnosis System. The Diagnosis System will then return a diagnosis back to the Mobile Device via the Secure Network. The diagnosis will provide a confidence score of the likelihood of diabetes in the patient.

II. METHODOLOGY

We collected anonymized TBI/PTSD Patient Data and develop NLP based Diabetes Predictive Engine. Differential Diagnostics or DDx is a method for determining the most likely disease that based on a set of patient symptoms. The basic theory of DDx is a *probabilistic measure* for estimating the likelihood of a specific diagnosis. In the case of Dibetes DDx, the measure would be

$$P(D) = \sum_{i}^{n} P(D/S_i)$$

Where

$P(D)$ = Probability of Diabetes in a patient
$P(D/S_i)$ = Probability of Diabetes in a patient, given Symptom *S*.

In this study in addition to the probability distribution method, we have used a machine learning approach (Support Vector Machine) to differentially diagnose Diabetes based on textual data collected based on the patient input.

III. DATA COLLECTION

A. Diabetes Check List

We collected the patient data from 25 people who were never pathologically tested for Diabetes. The Patient Data was divided into Training and Test data. The Automated DDx Systems was trained on one set of data and then tested on a blind set of data. There are two basic types of input for the Automated DDx: 1) Patient Intake Forms (Diabetes Check List) and 2) Free Text Patient Description.

[1] Contact Author - Shweta Sheel, MBBS, Gauhati Medical College Hospital, Assam, India; Email: shwetasheel@gmail.com; Veronica Heredia, MD; Email: vph@earthlink.net; Aman Kumar, BCL Technologies, San Jose, California; Email: amank@bcltechnologies.com.

Diabetes Checklist – Pilot Program

Client's Name: _____

Clinician: _____

Date: _____

Instruction to Client: Below is a list of common problems and complaints that are related to diabetic experience. Please read each one carefully, put an "X" in the box to indicate how much you have been bothered by that problem in the last 3 months.

No.	Response	Not at all (1)	A little bit (2)	Moderately (3)	Quite a bit (4)	Extremely (5)
1.	Do you have an urge for frequent urination than normal?					
2.	Do you have blurry vision recently?					
3.	Do you have increased hunger than normal?					
4.	Do you experience more fatigue than normal?					
5.	Do you experience dry mouth recently more than normal?					
6.	Do you feel more irritable than normal?					
7.	Do you experience unusual headaches?					
8.	Do you have had itchy skin?					
9.	Have you experienced Unusual Weight Changes?					
10.	Do you experience frequent infections such as frequent and persistent yeast infections in women, skin infections, urinary infections, or gum and mouth infections					
11.	Do you experience Sores, Cuts, and Bruises That Take a Long Time to Heal?					
12.	Do you experience Numbness or Tingling in the Hands or Feet?					
13.	Have you experienced Sexual Dysfunction lately?					
14.	Is your body shape an apple shape with thin arms and legs?					

Fig. 1. Diabetes Check List - Sample

B. Free Text Patient Description

In the absence of specific Diabetes Intake Forms, clinicians write down patient information in the form of free text. The figure below shows an example of free text input for a patient.

Mrs. <<anonymized>> is a 45 year old asymptomatic obese African American female who comes to your office for the first time for follow up of her DM. She was diagnosed with type 2 DM 6 months ago. The patient complains of burning sensation in his feet at night. On physical examination, you note decreased sensation in his toes and calluses on hos lateral fifth digits. She is taking glyburide 5 mg daily before breakfast. She is on no other prescription medications but she takes over the counter ibuprofen for knee pain. She follows the American Heart Association Diet. Her fasting blood sugar 2 months ago was 160.

Past medical history - positive for mild hypertension -130/85 to 140/90 mmHg, Hyperlipidemia: No Known drug allergies.

Family history- Positive for DM in her mother and older sister. Hypertension and coronary artery disease in her mother, father, older sister and younger brother.

Social history- Negative for smoking, alcohol or illicit drug abuse. Married with three adult children. Works as cashier at a local supermarket.

Fig. 2. Sample Case Presentation - Diabetes

The NLP based Diabetes Predictive Engine consists of 3 parts – Natural Language Parser, Semantic Role Labeler, and Support Vector Machine, as shown in the figure below:

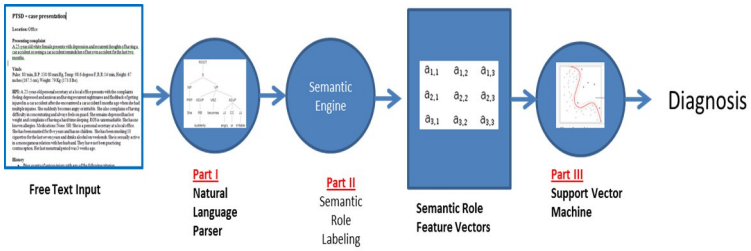

Fig. 3. Differential Diagnosis based on Text and Natural Language Processing

C. Natural Language Parser

The first step of the NLP Engine is to break the text into sentences and parse each sentence to find its grammatical structure and parts of speech. For instance the sentence:

> "The patient complains of burning sensation in his feet at night."

parses to the tree in the figure below:

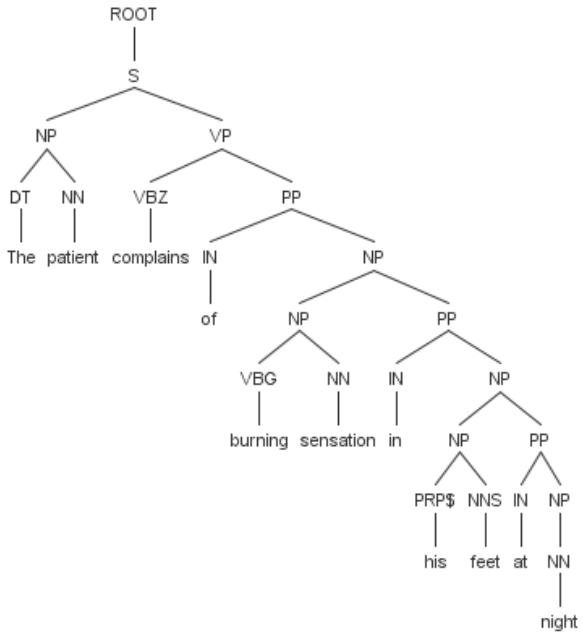

Fig. 4. Sample Parse Tree

Where:
DT: Determiner
S: Simple Declarative Clause
NP: Noun Phrase
NN: Noun, singular or mass
VP: Verb Phrase
VBZ: Verb, 3rd person singular present
PP: Prepositional Phrase
IN: Preposition
VBG: Verb, gerund/present participle
PRP$: Personal Pronoun
NNS: Noun plural

D. *Semantic Role Labeler*

This parse tree feeds into semantic role labeling module. This module (manual role labeling right now), along with a lexical ontology will semantically annotate Part-of-Speech (POS) tagged parse trees. In this case, the semantic roles are:

Agent: : <the patient>
Mood : <burning sensation>
Location: <feet>
Time: <night>

E. *Support Vector Machine*

These semantic roles form the Semantic Role Feature Vectors. Each document is converted into a matrix for feature vectors that are sent to a Support Vector Machine (SVM). We developed and trained the SVM to take Semantic Feature Vectors and predict Diabetes. We used Joachims SVM tool (http://svmlight.joachims.org/) that is a C implementation of Support Vector Machines. We have implemented two programs*: (1) svm_learn* , which takes a training file and creates a model based on it; and (2*) svm_classify* , which takes testing data and applies the model to it in order to classify it.

For differential diagnosis project, the input file includes the training examples. Each of the following lines represents one training example. The format of the training examples is given below.

<line> .=. <target> <feature>:<value>
<feature>:<value> ... <feature>:<value> # <info>
<target> .=. +1 | -1 | 0 | <float>
<feature> .=. <integer> | "qid"
<value> .=. <float>
<info> .=. <string>

During classification phase, the target value gives the class of the example. So, for a positive example (meaning the patient is diabetic) +1 as the target value, while -1 a negative example, meaning no diabetes, respectively.

IV. Experimental Set up

The acquisition of fully-anonymized 'Free Text Patient Description' and 'Diabetes Checklist' content for 25 adult men and women in the USA and India was done manually. These 25 people were not diagnosed with diabetes in the past. The textual data was fed into the preliminary diagnostic system. A phrase similarity repository is derived following Stanford's STRIDE ontology.

The preliminary DDDx system suggested if these people should be recommended for further screening of diabetes. The system provided a confidence score based on the term frequency matching of the features in the check list.

V. Evaluation

Based on the textual data entry of the 25 people the hand-simulated DDDx system recommended that 14 people should be further screened for Diabetes, meaning they were recommended to go for the following pathological tests under formal medical supervision:

- Fasting or pre-meal blood glucose
- Post-meal blood glucose measurement
- Hemoglobin A1C
- Dilated eye exam
- Comprehensive foot exam
- Urine test for microalbumin
- Blood pressure
- Weight
- Lipid control

After performing these pathological tests, we found that out of the 14 people that the DDDxA system recommended for further screening, 11 people tested positive for Type-2 Diabetes. Thus the preliminary DDDxA system recorded an accuracy of around **78.6%** accuracy. In other words, for the given sample size of 25 people, the DDDxA system could diagnose the Type-2 Diabetes in patients with a *precision* of 78.6%. The following table gives the preliminary results of this study.

Table I
Evaluation score of the Diabetes Differential Diagnosis Application System

Number of People Initially Screened (presumed healthy)	25
Number of people DDDxA system screened for potential Diabetes	14
Number of people confirmed to have Type-2 Diabetes after pathological tests under medical supervision	11
Accuracy of the DDDxA system	78.57%

VI. Conclusions and Future Work

In this study we developed a Diabetes Differential Diagnosis Application (DDDxA) System that takes textual data – 'Diabetes Check List' and 'Free Text Patient Data' as input and based on the textual and semantic analysis of the data recommends if a person should be recommended for further screening for Diabetes.

The DDDxA system has the ability to perform an initial diagnosis of the patient and provide initial treatment in the field away from a doctor's clinic. In addition, it can subsequently diagnose the patient based on response to treatment and help modify the diagnosis and treatment based on the patient's response to previous treatment.

For future work, we would like to expand the data set from 25 people to 100 people and evaluate the system. In addition, we would like to expand the 'check list' and 'free text patient data' to more features to have wider coverage and higher precision.

References

[1] Tucker ME. New AACE algorithm addresses all aspects of type 2 diabetes. Medscape Medical News [serial online]. April 23, 2013; Accessed May 1, 2013. Available at http://www.medscape.com/viewarticle/802954.

[2] Garber AJ, Abrahamson MJ, Barzilay JI, Blonde L, Bloomgarden ZT, Bush MA, et al. AACE Comprehensive Diabetes Management Algorithm 2013. Endocr Pract. Mar-Apr 2013;19(2):327-36.

[3] American Diabetes Association. Standards of medical care in diabetes--2012. *Diabetes Care*. Jan 2012;35 Suppl 1:S11-63l

[4] Keller DM. New EASD/ADA Position Paper Shifts Diabetes Treatment Goals. Medscape Medical News. Available at http://www.medscape.com/viewarticle/771989.

[5] Framenet: Frame Semantics Meets the Corpus. In LSA.Fillmore, C. and Baker, C. (2000)

[6] Automatic Labeling of Semantic Roles. Proc.of ACL.Gildea, D. and Jurafsky, D. (2000)

[7] Burges, Christopher J. C.; A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery 2:121–167, 1998