Interactive Medical Diagnosis System

Akash Saxena[1], Annu Sharma[2], Atanendu Sekhar Mandal[3]

[1]*Computer Science Department, Birla Institute of Technology and Science Pilani, India*

[2]*Computer Science Department, Birla Institute of Technology and Science Pilani, India*

[3]*Chief Scientist, IC Design Group, Central Electronics Engineering Research Institute, Pilani, India*

[1]f2010168@pilani.bits-pilani.ac.in

[2]f2010003@pilani.bits-pilani.ac.in

[3]atanu@ceeri.ernet.in

***Abstract*— The tasks of medical experts are becoming increasingly difficult with increasing population and consequently diseases. The people from rural area are at even greater difficulties as there is lack of availability of medical experts too. This project aims at developing an interactive system to provide a provisional diagnosis directly to the patients where there is lack of availability of doctors and physicians for one of the diseases – malaria, dengue or typhoid. It takes as input common natural utterances and outputs the best possible action to take, without any external help. The paper also proposes a methodology in which way the system can be used in real time scenarios.**

***Keywords*— Medical Diagnosis, Knowledge Base, Mutual Information, Natural Language Processing, Dengue, Typhoid, Malaria, Cloud Computing**

I. Introduction

The human-computer interaction plays a very vital role in bridging the gap between the patients, experts, medical officers or in general terms users of medical diagnosis systems. In recent years there have been a lot of researches in the field of developing Medical Diagnosis Systems for assisting experts and doctors in making decisions using multiple technologies such as Artificial Neural Networks[11], Artificial Intelligence[10] and Knowledge Based Systems[12,13]. However, most of the systems developed or are in the developmental process are majorly focused on for use by the experts or doctors for their assistance. Our system aims at providing the service to users from remote and low socio-economic background who do not have much access to doctors or medical hospitals without any outside assistance.

In today's world most of the people even those from rural region have mobile phone with the help of which a person can send a message in order to make a query to the system and get a provisional diagnosis. The system can be used to be deployed at any place and users can make query to it from different parts. In most of the scenarios diagnosis procedure usually starts with the patient chief complaints and the doctor learns more about the patient situation interactively during an interview, as well as by measuring some metrics such as blood pressure or the body temperature. The diagnosis is then determined by taking the whole available patients status into the account. Then depending on that, a suitable treatment is described, and the whole process might be iterated. In each iteration, the diagnosis might be reconfigured, refined, or even rejected [3]. Along with it also comes a series of problems in the process of diagnosis such as diverse symptoms generated by diverse causes should be considered. Epidemics also should be considered and the genetic factors too. In this paper we are considering a subset of those problems by limiting our dataset to 3 diseases namely dengue, typhoid and malaria along with their symptoms.

II. METHODOLOGY

Our approach presented aims at using some of the concepts defined above to come up with an innovative strategy for giving the directions to any patient who has a possibility of having any of the diseases mentioned. In our system, we assume that the patient will be providing the chief complaints in the form of a written text. The information thus given will be processed according to the algorithm given and steps explained to finally output the provisional diagnosis to be given to the patient. The diagnosis results include actions anywhere from taking rest, being alright to immediate hospitalization as would be the case.

A. Pre-processing

Pre-processed information present before the beginning of the procedure of identification of disease or suggestion of best action to be taken includes Knowledge base, symptom database and disease x symptom matrix.

1) *Knowledge Base:* KB considered here consists of a set of rules and inferences for the diseases under consideration and is constructed with the help of experts by interviewing them on these fields and medical text-books [1]. The disease x symptom (DxS) matrix is constructed and differential diagnosis is coded based on the rules provided by the experts. The DxS matrix represents the presence or absence of a particular symptom with respect to that particular disease. The mutual information gain presence and absence matrices are obtained after applying the formulas of MI and IG (presented later) to DxS matrix.

2) *Text Pre-Processing:* There are several methods of doing the pre-processing work to make the text available to us in the form of symptoms. We have used here the Natural Language Toolkit (NLTK) which is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language. This also forms the first step of our process where the raw text is processed through a series of supported NLP techniques by nltk, such as sentenizing, tokenization, spell corrector, keyword extraction and stemming represented in order of their execution [2]. Sentenizing carries breaking of the text into sentences. Tokenization feature breaks the sentence into tokens and the tokens thus obtained are spell corrected using a pre-written spell corrector in python [9]. We extract the keywords by using stop words removal and parsing the text using AIML(Artificial Intelligence Markup Language) [4] for keywords. The keywords are stemmed using the porter stemming algorithm [6] before starting the symptom analysing phase.

B. Mutual Information

In our scenario, after getting the DxS matrix representing the presence or absence of symptoms with respect to a disease from the knowledge base. We need to find the mutual relation between the two for carrying our further analysis. The mutual information of two random variables is a quantity that measures the mutual dependence of the two random variables. Formally, the mutual information of two discrete random variables X and Y can be defined as:

$$I(X;Y) = \sum_{y \in Y} p(y) \sum_{x \in X} p(x|y)\, log \frac{p(x|y)}{p(x)} =$$

$$\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad (1)$$

Where, p(x,y) is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y respectively. Mutual information is a measure of the inherent dependence expressed in the joint distribution of X and Y relative to the joint distribution of X and Y under the assumption of independence. Mutual information therefore measures dependence in the following sense: I(X; Y) = 0 if and only if X and Y are independent random variables [5]. Here the above formula is used to calculate the mutual information presence matrix for all the symptoms that are present in the DxS matrix. For the symptoms which are absent with respect to a disease the formula used is,

$$I(X;Y) = \sum_{y \in Y} p(\bar{y}) \sum_{x \in X} p(x|\bar{y})\, log \frac{p(x|\bar{y})}{p(x)} =$$

$$\sum_{x,y} p(x,\bar{y}) \log \frac{p(x,\bar{y})}{p(x)p(\bar{y})} \qquad (2)$$

Where, X is represented as the disease and Y as the symptom.

C. Diagnosis

A provisional diagnosis is one that is initially determined to be the diagnosis, except for the fact that all test results have not been received and/or analysed. When the full medical work-up has been completed, the provisional diagnosis is usually changed to a (definite) diagnosis. We start the provisional diagnosis given the set of symptoms provided by the patient. After obtaining the rules from the experts, we form our inferences for the order of examining (or in our case asking) the questions about their symptoms. For example, there is a 100% probability of having fever given the disease is malaria, typhoid or dengue; however the manner of occurrence may vary. Suppose a patient has entered fever as one of his symptom. Then the next step would be to analyse, what kind of fever is he suffering from? If not then the diagnosis to be made is based only on the given symptoms.

D. Algorithm Explanation

Input to the system: consists of the transcript of a patient similar to what is given by a patient to the doctor. The text obtained in this manner is full of un-useful information and backchannel.

Interaction between user and system: more knowledge about the problems of patient is obtained by the Q/A analysis.

Output of the system: consists of the most probable disease or best possible action to be taken by the patient.

Step 1: Pre-processing of the input text is done using python library NLTK. It does sentenizing, tokenization, stemming, Keyword extraction on the input text. The symptoms are extracted from the keywords obtained with the use of previously built database of symptoms.

Step 2: Disease matrix is constructed based on the analysed symptoms and their corresponding value in the knowledge base. We get the most crucial symptom (which has not been checked) of the most probable disease based on the matrix.

Step 3: Queries are made to the patient based on these crucial symptoms and the disease matrix is updated after sufficient analysis is done for identifying a disease or the patient has no more symptoms. The procedure is stopped and the appropriate action or disease is given as the output.

E. Use of Naïve Bayes Classifier

With the use of Mutual information theory we are able to provide the weightages to the different situations. Further on by the introduction of Naive-Bayes classifier a user is classified based on all his available attributes such as age, gender, location, social background and medical history. As a result a posterior probability of presence of disease is available before going into question and answer analysis. The classifier is trained on the available datasets of previously diagnosed patients. The dataset thus generated also adjusts the weights in the mutual information gain matrices as more and more patient data is inserted. Hence with time the system automatically learns from the datasets and which in turn will improve the accuracy of the whole system.

III. Experiments and Results

The experiments were performed based on the inputs given by the experts and the output of the system that is, the provisional diagnosis was returned to the experts for administering their validity. For the analysis many experiments were performed to determine the correctness and accuracy of the program. However, here we only show some different varieties and complexities of inputs on which the system was tried. The snapshots from different inputs are shown below:

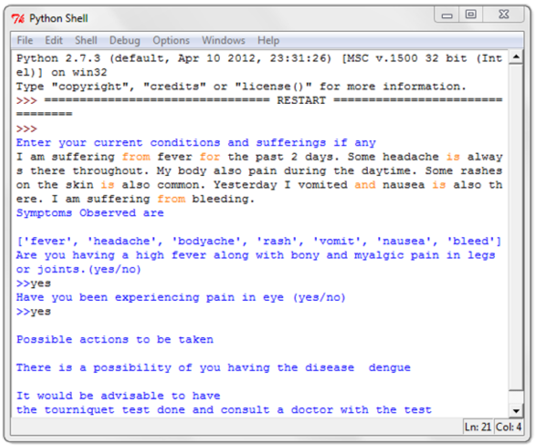

Fig 1 Person showing all the signs and symptoms of dengue.

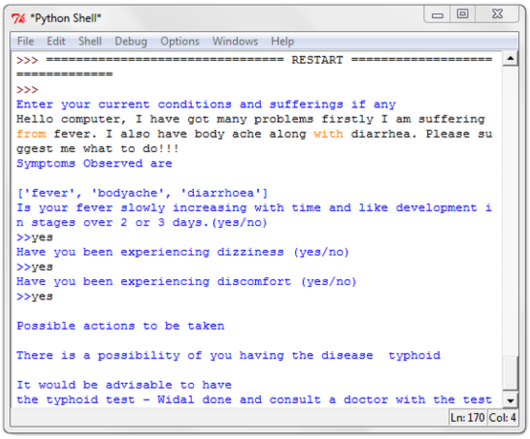

Fig. 2 Person showing some of the symptoms of typhoid as chief complaint, rest is asked in the form of question for the confirmation.

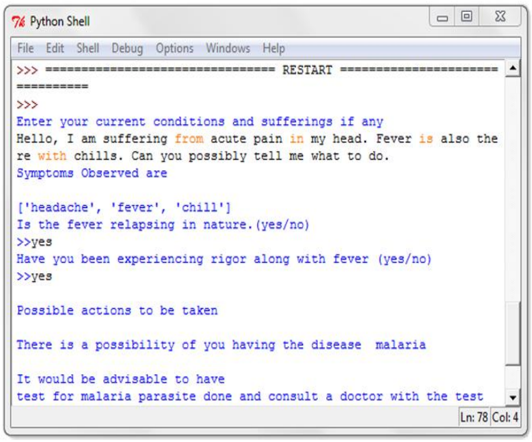

Fig. 3 Person showing some of the symptoms of malaria and other common symptoms as chief complaint, rest is asked in the form of question for the confirmation.

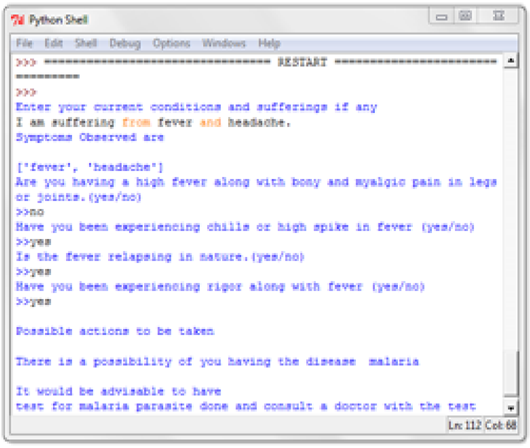

Fig. 4 Only very few and common symptoms are provided by the patient as chief complaint. The system changes its decision based on the inputs given and considering possibilities.

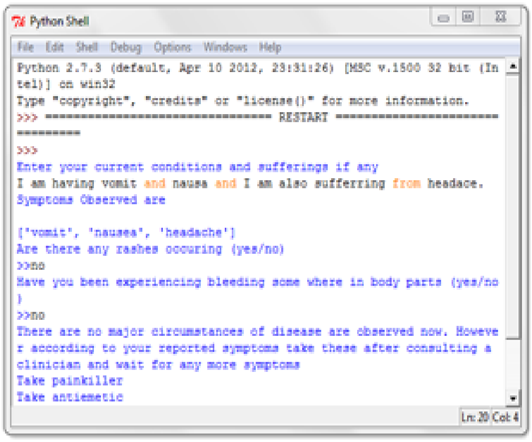

Fig. 5 A demo example to illustrate that system can even correct some minor mistakes in spelling. As it might be dealing with unknown users in the future. It also depicts the different provisional diagnosis to be given in case of only such symptoms.

As stated above the outcomes of the system were presented to the experts for their assessment. The experts finally agreed to the manner and final results that the system showed. According to the experts the scope of the system can further be improved if it can be tested on real datasets which can be obtained from a hospital or medical centre.

IV. WORKING OF THE MODEL

A proposed methodology for deploying the system in real time scenarios. Tools and techniques used will be Google App Engine [8] for hosting and running the app on cloud so that it's easily accessible and reachable, Twilio app [7] for receiving queries communicating it to the cloud for processing and sending it back to the user. The detailed model working is as follows: The user sends the information in the form of a speech along with an image or any textual matter to the number given as default by the twilio account; the app receives it using the TwiML and the twilio Rest API. The message received is sent to the app hosted on the Google cloud for processing using the same APIs. The app on cloud or earlier referred to as the system then processes the information and sends back the diagnosis or the question to be asked from the user. In similar manner, the twilio app passes that onto the user in the form of speech or text using the text-speech api and the data about the conversation is stored in the cookie for future reference. In this way a two way communication can be maintained between the user and the system without any external help.

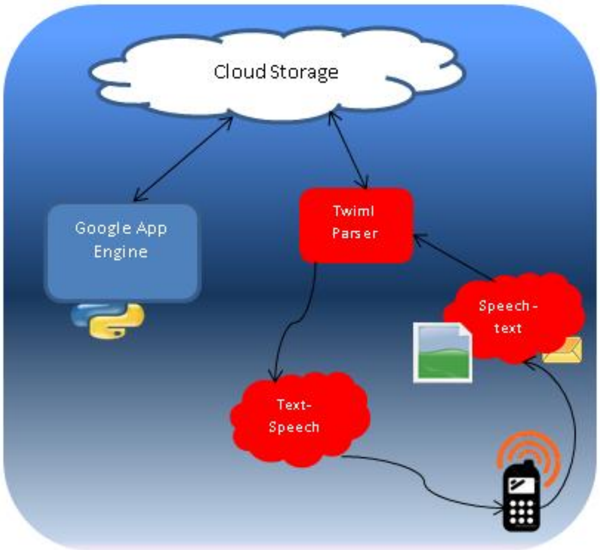

Fig. 6 Figure showing the overall picture about the interactions between users the api and python code run on Google app engine.

V. CONCLUSION

In this system we tried to provide a provisional diagnosis, but without concerning about how to calculate the best relation between the symptom and the disease. Actually, with the help of knowledge extracted from experts and the Mutual Information we were able find out the relation. This means that instead of a patient consulting the doctor for mild cases of malaria, typhoid and dengue complications, the patient can receive treatment from the house through the Medical Diagnosis system. However, in future with the help of training set of patients suffering or non-suffering from the disease but having symptoms close to it. We would be able to refine the system's accuracy in providing the diagnosis.

VI. FUTURE WORK

As for future work, one may concentrate on extending the database to incorporate more diseases which are common. Advancement in the field of natural language processing algorithms for sentiment analysis and semantic similarity will help in better understanding of the patient's concern. In the working model, with the help of hospitals and doctors the system can be modified to contact them directly in case of any adverse situation of a patient. Also with the advancement of field of Computer Vision and sensors technology the image analysis and sensor data will reveal more information about the patient's health and status then the mere textual information is able to provide.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable comments which helped to improve the paper. The authors also thank the experts for sharing their knowledge and spending their time without which this paper could not have been completed.

REFERENCES

[1] Longo, Fauci, Kasper, Hauser, Jameson, Loscalzo *Harrison_'s Principles of Internal Medicine*, 18th Edition, McGraw-Hill, 2012, Vol 1.

[2] Bird, Steven, Edward Loper and Ewan Klein (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

[3] S. Moein, S. A. Monadjemi, and P. Moallem (2008). "A Novel Fuzzy-Neural Based Medical," *International Journal of Biological and Life Sciences*, pp. 146-150, 2008

[4] A.L.I.C.E. homepage on AIML: Artificial Intelligence Markup Language. [Online]. Available: http://www.alicebot.org/aiml.html

[5] Mutual Information homepage on Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Mutual_information

[6] M. Porter (2006) The Porter Stemming Algorithm homepage on Porter Stemming Algorithm. [Online]. Available: http://tartarus.org/~martin/PorterStemmer/index.html

[7] The Twilio Website. [Online]. Available: https://www.twilio.com

[8] Google App Engine - Google Developers website. [Online]. Available: https://developers.google.com/appengine/

[9] P. Morvig How to Write a Spelling Corrector homepage on How to Write a Spelling Corrector. [Online]. Available: http://norvig.com/spell-correct.html

[10] Smita Sushil Sikchi, Sushil Sikchi, and M. S. Ali, "Artificial Intelligence in Medical Diagnosis," in *International Journal of Applied Engineering Research*, 2012

[11] Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara ,Josef Havel and Aleš Hampl "Artificial neural networks in medical diagnosis," in *Journal of APPLIED BIOMEDICINE*, 2013

[12] P.Santosh Kumar Patra, Dipti Prava Sahu and Indrajit Mandal "An Expert System for Diagnosis of Human Diseases," in *International Journal of Computer Applications*, 2010

[13] S.A. Fatumo, Emmanuel Adetiba and J.O.Onaolapo "Implementation of XpertMalTyph: An Expert System for Medical Diagnosis of the Complications of Malaria and Typhoid," in *OSR Journal of Computer Engineering*, 2013