International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 07 | July-2016 www.irjet.net p-ISSN: 2395-0072

DISEASE INFERENCE SYSTEM ON THE BASIS OF HEALTH RELATED QUESTIONS VIA LEARNING

Shaima P.[1], Radhika K.T.[2]

[1]*M Tech Student, Dept. of Computer science& Engineering, M.E.A Engineering College, Perinthalmanna, Kerala*
[2]*Assistant Professor, Dept. of Computer Science& Engineering, M.E.A Engineering college, Perinthalmanna, Kerala*

Abstract - *Better health is central to human happiness and well being. But diseases are one of the increasing subject. The health seeker have many online and offline methods to get the information requested by them. The researchers are encouraged by the advancement in computer technology and machine learning techniques are used to develop software to assist doctors in making decision without necessitating the direct consultation with the specialists. Automatic disease inference is significance to overcome the difficulty of online health seeker. Here propose a novel deep learning scheme to infer the possible disease given the question of health seekers. In this work first analyze the information needs of health seekers in terms of question and then select those that ask for possible diseases of their manifested symptoms for further analytic. Then user will search for their needs as query. Next preprocesses the query to find the medical attributes. Then the preprocessed attributes to identify the corresponding disease concept.*

Key Words: **Health seeker, Disease inference, Deep learning, Symptom, Medical attribute, Query.**

1. INTRODUCTION

Medical information is accessible from diverse source including the general web, social media, journal articles, and hospital records. Users may patients and their families, researchers, practitioners and clinicians. Challenges in medical information retrieval include variations in the format, reliability, quality of biomedical and medical information. Online health information includes both medical resources and patient community connections. They continues to play an important role in patient education and self-care. Results from a national consumer survey conducted by Makovsky Health and Kelton show the average U.S. consumer spends nearly 52 hours looking for health information on the internet annually, and visits the doctor three times. Physicians remain a key influencer sparking online health research, Americans are most likely to visit a pharma-sponsored website after receiving a diagnosis from their physician. These findings underscore the importance of accuracy and accessibility of online health information as a springboard for patient-physician dialogue and peer support.

The current prevailing online health resources can be roughly categorized into two categories. One is the reputable portals run by official sectors, renowned organizations or other professional health providers. They are disseminating up-to-date health information by releasing the most accurate, well-structured, and formally presented health knowledge on various topics. WebMD and MedlinePlus are the typical examples. The other category is the community-based health services, such as HealthTap and HaoDF. They offer interactive platforms, where health seekers can anonymously ask health-oriented questions while doctors provide the knowledgeable and trustworthy answers. However, the community based health services have several intrinsic limitations. First of all, it is very time consuming for health seekers to get their posted questions resolved. The time could vary from hours to days. Second, doctors are having to cope with an ever-expanding workload, which leads to decreased enthusiasm and efficiency.

Health seekers frequently ask for supplemental information of their diagnosed disease, preventive information of their concerned diseases or possible diseases of their manifested signals. Table-1 shows the categories of health seeker needs with example.

Table-1 : Categorization of health seeker needs

Categories	Question Example
Disease diagnosed, ask for supplemental information	I have been diagnosed with polyarteritis nodosa and I am currently taking 32mg of steroids and 125mg of

	cyclophosphamide. Is this the correct medicine ?
Disease undiagnosed, ask for possible diseases of their manifested signals	What disease or illness should I look into if I feel tired, sleepy all the times, muscular and joint sored, decrease in memory and concentration ?
Currently healthy, ask for prevention	I'm a healthy and active 38 year-old Asian male whose non-smoking mother died from lung cancer 4 years ago. Am I at greater risk ? How can I pre-empt it ?

2. RELATED WORK

There are lots of classification methods such as KNN(K Nearest Neighbor)[1], Naive bayes classifier[2],SVM[3].

KNN is a question classification method based on vector space model. Assume *x* is a question, and its vector model is x=(x_1 ,x_2....x_n). Each dimension of question vector *x* corresponds to each word of question representation, also known as attribute. C_i= (x_1^i, x_2^i..., x_n^i) is a question category with class identifier containing question x_1^i, x_2^i..., x_n^i. Let there be *m* question training classification questions . C_1, . C_2,..... . C_m, the classification process of KNN is as follows: for a given test question *x*, in the training question set of all categories C_1, . C_2,..... . C_m, , the similarity between two questions can be used to find *k* =*k* >1 nearest training questions. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong independence Assumptions. A naive Bayes classifier follows conditional independence since it assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature, given the class variable. Thus terms are given a weight value which is independent of its position and presence of other terms. Naive Bayes classifier is trained by set of labeled training examples.

Methods for extracting medical entities are Support Vector Machine[4], Temporal Abstraction[5], A convolutional approach Framework[6]. SVM is a machine learning method that is widely used in many NLP tasks such as chunking, POS, and NER. Essentially, it constructs a binary classifier using labeled training samples. Given a set of training samples, the SVM (Support Vector machine)training phrase tries to find the optimal hyperplane, which maximizes the distance of training sample nearest to it . SVM takes an input as a vector and maps it into a feature space using a kernel function. The temporal abstraction framework for classifying the patient's time-series data based on temporal abstractions. The temporal abstraction framework deals with STF-Mine algorithm automatically mines discriminative temporal abstraction patterns from the data and uses them to learn a classification model. Convolutional Approach is deals with a Nonnegative Matrix Factorization (NMF) based framework for open-ended temporal pattern discovery over large collections of clinical records. The framework can mine common as well as individual shift-invariant temporal patterns from heterogeneous events over different patient groups, Temporal Pattern Discovery (TPD) for EHR data, which aims at finding temporal patterns of one or more groups of patients.

Inference methods are multi switch Transductive SVM[7], Deterministic Annealing Semi-supervised SVM. Transductive SVM is well known for linear semi-supervised classification on large and sparse data sets. Transductive SVM appends an additional term in the SVM objective function whose role is to drive the classification hyper plane towards low data density regions. It is able to alleviate the problem of local minimum in the TSVM optimization procedure while also being computationally attractive. Now present a new algorithm based on deterministic annealing that can potentially overcome the transductive SVM loss function problem while also being computationally very attractive for large scale applications. Deterministic Annealing is an established tool used for combinatorial optimization that approaches the problem from information theoretic principles.

3. PROPOSED SYSTEM

Main objective of the system is to improve efficiency. This scheme builds a novel deep learning model. First mines the latent medical signatures from the health related reviews. They are the compact patterns of inter-dependent medical terminologies or raw features. This information is used for deep learning. The raw features and signatures respectively serve as input nodes in one layer and hidden nodes in the subsequent layer. The second learns the inter-relations between these two layers via pre-training. The hidden nodes are viewed as raw features for more

abstract signature mining. This scheme builds a sparsely connected deep learning architecture with three hidden layers. This model is generalizable and scalable. Fine-tuning with a small set of labeled disease samples fits our model to specific disease inference. The number of hidden nodes in each layer of our model is automatically determined than conventional deep learning algorithms and the connections between two adjacent layers are sparse, which make it faster. The proposed method consist of following steps:

1. Classification of health related questions.
2. Extraction of medical attributes.
3. Disease inference.

3.1 Support vector machines

Support Vector Machines (SVM)[2] are linear functions of the form $f(x) = w \bullet x + b$, where $w \bullet x$ is the inner product between the weight vector w and the input vector x . The SVM can be used as a classifier by setting the class to 1 if $f(x) > 0$ and to -1 otherwise. The main idea of SVM is to select a hyperplane that separates the positive and negative examples while maximizing the minimum margin.

3.2 Metamap

The Metamap become established as one of the premier applications for the identification of Metathesaurus concepts in biomedical text[8]. Because of ongoing interaction with user community, MetaMap arose in the context of an effort to improve biomedical text retrieval, specifically the retrieval of MEDLINE or PubMed . It provided a link between the text of biomedical literature and the knowledge, including synonymy relationships, embedded in the Metathesaurus. Early MetaMap development was guided by linguistic principles which provided both a rigorous foundation and a flexible architecture in which to explore mapping strategies and their applications. A system diagram showing MetaMap processing is shown in figure 1. Input text undergoes a lexical/ syntactic analysis consisting of:

- Tokenization, sentence boundary determination and acronym/abbreviation identification.
- Part-of-speech tagging
- Lexical lookup of input words in the SPECIALIST lexicon
- A final syntactic analysis consisting of a shallow parse in which phrases and their lexical heads are identified by the SPECIALIST minimal commitment parser. Each phrase found by this analysis is further analyzed by the following processes:
- Variant generation, in which variants of all phrase words are determined.
- Candidate identification, in which intermediate results consisting of Metathesaurus strings, called candidates, matching some phrase text are computed and evaluated as to how well they match the input text.
- Mapping construction, in which candidates found in the previous step are combined and evaluated to produce a final result that best matches the phrase text.
- Word sense disambiguation (WSD), in which mappings involving concepts that are semantically consistent with surrounding text are favored.

The evaluation performed on both the candidates and the final mappings is a linear combination of four linguistically inspired measures: centrality, variation, coverage and cohesiveness. The evaluation process begins by focusing on the association or mapping of input text words to words of the candidates. Centrality, the simplest of the measures, is a Boolean value which is one if the linguistic head of the input text is associated with any of the candidate words. The variation measure is the average of the variation between all text words and their matching candidate words, if any. Coverage and cohesiveness measure how much of the input text is involved in the mapping and in how many chunks of contiguous text . The four measures are combined linearly giving coverage and cohesiveness twice the weight of centrality and variation.

3.3 Sparsely Connected Deep Learning

The sparsely connected deep learning model has L layers with $d_l (1 \leq l \leq L)$ nodes in each layer[9][10]. To be more specific, the first layer contains the input n-dimension raw features and the L-th layers denotes the output disease types, while the intermediate layers are hidden layers, which are unseen from the data. Unlike general deep learning architectures, in this work, nodes in the higher layer are the signatures of and connect to the nodes in its adjacent lower layer, rather than fully connected. The three hidden layers were constructed incrementally, alternating between subgraph mining and pre-training.

Initially regarded the learning model with only one hidden layer. Each node in the hidden layer is corresponding to a signature obtained via dense subgraph mining from a large graph. Figure 2 shows the illustrative process of sparsely connected deep learning construction. It incrementally added hidden layers until it satisfies the predefined convergence criterion. Figure-3 shows the Block diagram of proposed work.

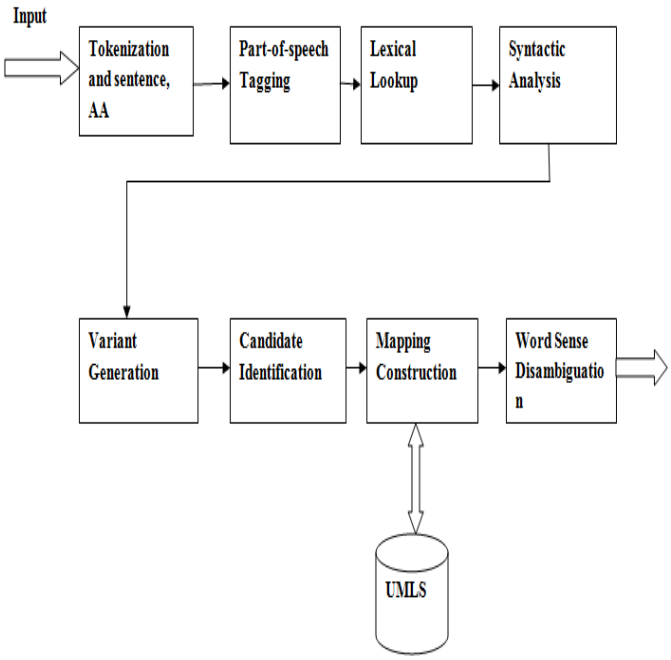

Fig -1: Metamap system diagram

4. EXPERIMENTS

4.1 Dataset

Disease concepts are used in this experiment. Collect disease concepts from WebMD is an American corporation known primarily as an online publisher of news and information pertaining to human health and well-being. WebMD is best known as a health information services website, which publishes content regarding health and health care topics, including a symptom checklist, pharmacy information, drugs information, blogs of physicians with specific topics, and providing a place to store personal medical information. They cover wide range of diseases, including endocrine, urinary, neurological and other aspects. With these disease concepts as queries, can crawle more than 220 thousand community generated QA pairs from HealthTap.

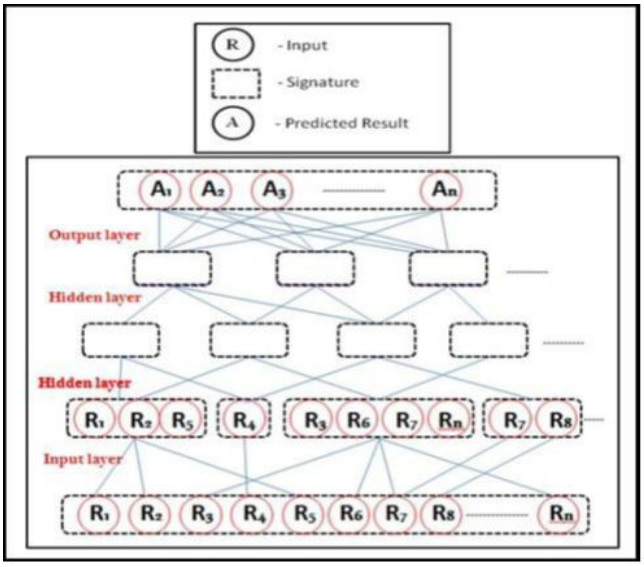

Fig-2: Overview of the process of sparsely connected deep learning construction.

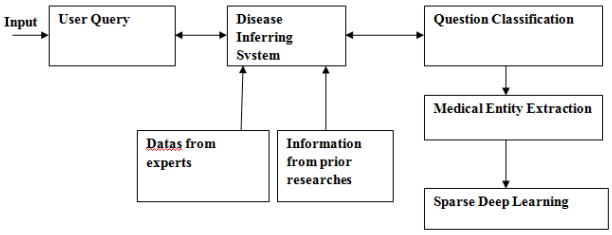

Fig-3: Block diagram of proposed work

4.2 Health information needs Analytic

First randomly selected Question answer pairs and they are read and understood. Then Summarized the health information needs into three categories such as disease diagnosed, disease undiagnosed and asking for preventive knowledge.

4.3 Extraction

The medical attributes are identified with Metamap tool. We get terminologies after removal of low frequency and normalization.

4.4 Inference

The deep learning model contains many layers including input and output layers. The nodes in the input layer represents raw features extracted and the output layer denotes the inference. This inference is the resultant disease to the user.

5. CONCLUSIONS

Medical diagnosis has become highly attributed with the development of technology lately. Furthermore the computer and communication tools have improved medical practice implementation to a greater extent. In this paper, we proposed a method to detect disease inference from queries given by health seeker and with the help of data set collected from various source. First performed analysis of the health seeker needs. This provides the insights of community based health services. Then presented a sparsely connected deep learning scheme that is able to infer the possible diseases given the question of health seekers. This scheme is constructed via alternative signature mining and pre-training in an incremental way. It permits unsupervised feature learning. Therefore, it is generalizable and scalable as compared to previous disease inference using other learning approaches. Learning architectures are densely connected and the node number in each hidden layers are tediously adjusted. In contract, The proposed model is sparsely connected with improved learning efficiency, and the number of hidden nodes are automatically determined and they improve accuracy.

REFERENCES

[1] Chen Caixian, Han Huijian, Liu Zheng," KNN question classification method based on Apriori algorithm" *computer modelling & new technologies*, pp.371-379, 2014.

[2] Rishika Yadav, Megha Mishra," Question Classification Using Naïve Bayes Machine Learning Approach" *International Journal of Engineering and Innovative Technology (IJEIT),* Vol 2, February 2013

[3] Dell Zhang, Wee Sun Lee," Question Classification using Support Vector Machines" *SIGIR'03*, August 1, 2003.

[4] S. Doan and H. Xu, "Recognizing medication related entities in hospital discharge summaries using support vector machine*," in Proc. Int. Conf. Comput. Linguistics*, 2010, pp. 259–266. ., 2008, pp. 29–33

[5] I. Batal, L. Sacchi, R. Bellazzi, and M. Hauskrecht, "A temporal abstraction framework for classifying clinical temporal data," *in Proc. Amer. Med. Informat. Assoc,2008,pp.29-33.*

[6] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach," in *Proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining,* 2012, pp. 453–461.

[7] V. Sindhwani and S. S. Keerthi, "Large scale semi-supervised linear SVMs," in Proc. 29th Int. ACM SIGIR Conf. Res. Develop. Inf.Retrieval, 2006, pp. 477–484.

[8] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: Historical perspective and recent advances," J. Amer. Med. Informat. Assoc., vol. 17, no. 3, pp. 229–236, 2010.

[9] Y. Bengio, "Learning deep architectures for ai," Found. Trends Mach. Learn., vol. 2, no. 2, pp. 1–127, 2009.

[10] Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan, Bo Zhang, Tat-Seng Chua," Disease Inference from Health-Related Questions via Sparse Deep Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no.8, August 2015 .