iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** XI **Month of publication:** November 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Survey on Disease Inference to Limit Medical Verbiage Discrepancy between Online Health Seeker and Provider

R.Meena Gomathi[1], Dr.S.Miruna Joe Amali[2]
[1]*PG student,* [2]*Associate professor, Department of CSE,*
K.L.N.College of Engineering, Madurai, Tamilnadu, India

Abstract-Bridging the gap between what online well-being seekers with unusual signs requisite and what busy human doctors with biased proficiency can offer is of greater importance. Online health care forums have been assisting well-being condition monitoring, illness modelling and validation of medical treatment by medical text mining. Precisely and competently concluding the diseases is non-trivial, especially for community-based health services due to the lexis gap, inadequate information, interrelated medical concepts, and incomplete high quality training samples. In this survey, the information needs of health seekers in terms of with their manifested symptoms are studied. An extensive learning structure is used to infer the diseases given the queries of health seekers. This scheme has two key components which first globally mines the discriminant medical signatures from raw features. And then it estimates the raw features and their signatures as input nodes in one layer and hidden nodes in the subsequent layers. The inter-relations between these two layers are found. All-encompassing trials on a real-world dataset labelled by online doctors show the noteworthy performance gains.

Keywords: Online Health Care Forums, Signatures, Medical Text Mining, Illness Modelling, Lexis Gap.

I. INTRODUCTION

Medical text mining is a method of discovering the hidden patterns in the data sets of the medical ground. The principal online health services suggest an interactive standard, where health seekers can ask health-oriented questions while clinicians provide the knowledgeable and trustworthy answers. The volume of health centre summaries written in natural language is rapidly increasing; physicians need a tool to automatically extract information about diseases/treatments. The main delinquent in extracting medical information is that physicians use variant words to label the same disease or treatment. In order to help physicians interpret and share disease facts in clinic records, we need to reliably and effectively identify and normalize the medical lexicons. This leads to a surge in distinguishing corpus sentient jargons from the raw details of the user posed queries. These are further refined for automatic disorder discovery.A review reports that an user spends adjacent to 52 hours annually online to discover health facts, while only visits the physicians three times per year. These verdicts have intensified the prominence of online health assets as catalysts to facilitate patient-doctor communication.The existing online health resources can be coarsely characterized into two kinds. One is the reliable portals run by official sectors, popular organizations, or other expert health providers. They are publicizing up-to-date health facts by discharging the most accurate, well organized, and properly offered health acquaintance on various subjects. WebMD and Medline Plus are the typical examples. The other category is the community-based health services, such as HealthTap and HaoDF.Community based fitness services have numerous inherent limitations. It is very time consuming for health inquirers to get their posted queries resolved. The period could diverge from hours to days. The physicians have to handle with an ever-expanding workload, which leads to decreased interest and efficacy. There is a lexis disparity among health seekers and health care provider, in order to amortize the gap local mining and global learning approaches are used. Generally, the community laid content, may not be directly available due to the medical verbose gap. Users do not share similar vocabulary. For e.g. HealthTap, is a question answering site for users to request fitness related queries. The queries are written by our own words. The similar question may be described in substantially multiple ways by two distinct health seekers of diverse background. The answer provided by the health care professionals may contain expression with different possible connotations, and non-standardized terms. The tags used generally may not be medical terminologies. For e.g., “heart attack” and “myocardial disorder” is similar medical terms referred by multiple experts. Users had encountered big challenges in reusing the archived files due to the illogicality between their quest terms and those gathered medical records. Automatically coding of medical records is immensely desired using standardized terminologies. This survey paper summarizes methods to normalize medical

verbose in order to curtail the gap among health inquirer and provider.

II. LITERATURE REVIEW

L. Nie et al., [1] have presented their effort to inhibit the terminology gap among health inquirers and providers which deferred the cross system inter-operability. Most of the existing health providers organize and code the medical archives by hand. There is an emergent concern to progress automated approaches for medical jargon assignment. Local mining and global learning methods are conjointly exploited. Local mining targets to locally code the health registers by mining the medical concepts from discrete record and then mapping them to lexicons based on the exterior legitimate terminologies.

A tri step framework, which includes noun phrase mining, medical model recognition and normalization, is established. Global learning learns mislaid key perceptions and disseminates exact jargons in the midst of underlying connected records over a large collection. The prevailing practices can be characterised into two kinds: rule based and machine learning tactics. Rule-based approaches play a vital part in medical terms assignments. It builds a novel global learning model to collaboratively enhance the local coding results. This guarantees the relevance probability function in continuous and smooth semantic space. The approach is mainly related to empirical loss function which forces the relevance probability. It lacks in learning about the heterogeneous cues. They ascertain and build operative rules by building robust uses of the syntactic, semantic and accurate aspects of natural linguistics. Machine learning approaches build inference prototypes from medical facts with identified observations and then relate the accomplished models to hidden data for terminology forecast. The whole process of the proposed method is unsupervised and it holds potential to handle large-scale data. M. Wang et al., [2] anticipated a methodology that inevitably regulates which sort of media data should be added for a textual response. It spontaneously gathers facts from the net to augment the answer. By processing an enormous set of QA duos and accumulating them to a group, this methodology can qualify a unique multimedia question answering (MMQA) approach as consumers can discover multimedia responses by equating their questions with those in the group. For a given QA pair, scheme proposed in [2] first predicts which type of medium is appropriate for enriching the original textual answer. Following that, it automatically generates a query based on the QA knowledge and then performs multimedia search with the query. Proposed diverse relevance ranking scheme for social image search, which is able to simultaneously take relevance and diversity into account. It leverages both visual information of images and the semantic information of tags.

Berlanga, Rafae et al., [3] intended an approach to produce word-concept likelihoods from a Knowledge Base (KB) that aids as a foundation for numerous text mining jobs which not only takes into account the core patterns within the descriptions enclosed in the Knowledge Base (KB) but also those in texts presented from huge unlabeled corpora such as MEDLINE. This system attains a higher degree of precision than other state-of-the-art methods when estimated on the MSH WSD data set. Ben Abacha et al., [4] extended a practice that contracts with different types of queries, including questions with more than one expected answer and more than one focus. The Question Answering (QA) scheme is used to provide precise and quick answers to user questions from a pool of documents or a database. This kind of IR system is very important for the growth of digital information. This paper addresses the problem of QA in the medical domain. A semantic approach to QA based on (i) Natural Language Processing techniques, it allows a deep analysis of medical questions and documents and (ii) semantic Web technologies at both representation and interrogation levels was proposed. Semantic Question-Answering System is called as MEANS is based on semantic search and query relaxation. The overall system performance on real world data set which includes queries and responses extracted from forums such as MEDLINE, WebMD, Health Tap.It allows a profound inspection for queries and masses by means of diverse information extraction approaches. The outline uses both vast scrutiny of question and documents in order to abstract information. Identify and mine the well-being entities (e.g. diseases, drugs, symptoms). Difficultly level lies in determining the classification of semantic relations between these entities (e.g. treats, prevents, causes).

Frunza, Oana; Inkpen, Diana; Tran, Thomas [5] in the year 2011 has put forward a technique to recognize and distribute healthcare info and comprehend the semantic associations that arise between syndromes and treatments. It involves two responsibilities (i) routinely recognizing sentences available in medical summaries (Medline) which includes indication about syndromes and treatments, and relay to semantic relationships that occur amid diseases and treatments. (ii)The second task is committed on three semantic relations: Cure, Avert, and Consequence. The noun-phrases, verb-phrases, and biomedical concepts are acknowledged in the verdicts. Unified Medical Language system (UMLS) idea illustrations are an evidence source which encompasses a Meta thesaurus, a semantic organization, and the expert lexis for biomedical field.

F. Wang, N. Lee, J. Hu et al.,[6] proposed novel temporal event matrix demonstration and knowledge structure in aggregation with an

in-depth justification on equally synthetic and real world datasets. This framework enables the depiction, abstraction, and mining of high order event structure and relationships within single and multiple event sequences. Temporal Event Signature Mining involves two processes, i) One-Sided Convolution ii) β-divergence. The diverse event sequences are mapped to a geometric image by training events as a structured spatial-temporal shape process. It optimizes the performance of large-scale incremental learning of group-specific temporal event signatures. Then validates the framework on synthetic data and on an electronic health record dataset.

III. RELATED WORK

L. Nie et al., [1] has used Concept Entropy Impurity (CEI) methodology to relatively detect and normalize the health concepts locally, and construct a corpus responsive vocabulary with the assistance of exterior understanding. The local mining consists of the noun phrase mining, medical ideas detection and normalization. The global mining comprises of inter expert association; inter terminology relationship, probabilistic hyper graph creation. Global learning model is fabricated to collaboratively augment the local coding outcomes. This model flawlessly assimilates numerous heterogeneous information cues.

The performance evaluation of the two methods, viz. local and global mining was studied. The metrics of evaluation included accuracy, time and quality of work. The accuracy in efficiently analysing the health inquirers need was comparatively low in local mining when equated with global mining technique. The time to retrieve the remedies/responses for the given user posted queries was ranked to be immensely high in the global mining approach. Likewise the quality of outcome from the designed methods seem to have drastic improvements when a combination of both local and global mining methods are used. The methods proposed by L. Nie et al., [1] is completely indexed and thus the retrieval time is faster. In case of resource insufficiency the Query and the Question will be left in pending state till an expert arrives. When professionals go through the query, the responses not only dispatch to the wellness seekers and also update the local mining database for forthcoming instantaneous recovery to the related request from other users. The global learning approach is developed to compensate for the inadequacy of local coding method.

M. Wang et al., [2] has proposed three practices. For a question, retrieve question answer pair from the available question answering sites database dynamically and select an answer medium to enrich the textual answer. Then generate a query for the multimedia search, resulting data are undergoes duplicate elimination and irrelevant data removal by the help of graph based re ranking. Finally present the answer that contains textual data, images and videos. 40% of the questions in forums have mostly one manually labelled tag, because the questions are not in brief form. This is also causes a problem for the users who wishes to get an response. These tags are irrelevant and this incompleteness of question tags blocks all the tag-based manipulations, such as feeds for topic-followers, ontological knowledge organization, and other basic statistics. This paper presents a novel method that is able to comprehensively learn descriptive tags for each question.They 3 practices include

Answer medium selection- Given a QA pairs; it foretells whether the textual reply ought to be complemented with media evidence, and which kind of media files must be added.

Question generation for multimedia exploration-Given a QA pair, this component extracts three queries from the inquest, the response, and the QA duos, compatibly. The most informative question will be selected by a three-class ordering model.

Multimedia data selection and exhibition- Based on the produced queries, we gather image and video data with multimedia exploration engines.

The gathering precisions for inquiry selection with diverse features such as POS histogram and search performance prediction (SPP) are used. By processing a large set of QA pairs and adding them to a pool, this approach can enable a novel multimedia question answering (MMQA) approach as users can find multimedia answers by matching their questions with those in the pool. The existing system lack of diversity of the generated media data.

Frunza, Oana; Inkpen, Diana; Tran, Thomas [5] have suggested a mode to disseminate healthcare facts and understand the semantic associations that exist between diseases and treatments. As a sorting procedure, representative models such as: decision-based models probabilistic models (Naive Bayes (NB) and Complement Naive Bayes (CNB), adaptive learning (AdaBoost), a linear classifier (SVM) is utilized. This ML-based practice is used for constructing an application that is skilful in recognizing and splitting healthcare information swiftly. This obtains a reliable outcome that could be integrated in an application to be used in the health care domain.

Fleuren et al.,[7] The experimental data for biomedical research has been proliferated these days. The text mining tools are improved in quality and are playing a vital role in many research questions categorization ranging from de novo drug target discovery to

enhanced biological interpretation. This paper introduces techniques that are driven from the ABC principle (The ABC principle is used to get the relationship between the two terms with the help of the other term. And Without the third term the first two terms are not related directly to each other). This paper main proposes methods that are used for a text mining and an overview of the text mining tools. The scheme used has the advantage that indirect links between concepts become apparent, which can give insight into for instance new relations between genes or previously unknown gene disease associations. It is also applied to solve actual research questions. The only drawback is that these methods are trained for detection of specific relation-ships on a training set and are thus limited by the availability and quality of the training data

Liqiang Nie et al.,[8] has put forward a model to analyse the user posted questions which excerpts corpus recognizant medical lexicons from those raw features. This step is termed to be called as "signature mining". A deep learning technique is used, which at preliminary level matches the symptoms discretely with a disease. It results in inferring various diseases for the manifested signals. Correspondingly at descendant levels, combination of multiple symptoms is checked in order to predict the disease. This augments the level of accuracy in inferring the disease.

IV. RESULT ANALYSIS

The approach proposed by Liqiang Nie et al., [8] is showed in the table 1. Machine learning approaches form inference prototypes from medical information with well-known comments and then relate the trained models to hidden information for terminology forecast. This paper aids from the volume of unstructured community generated data and it is proficient to handle various kinds of syndromes successfully. It explores and classifies the information requirements of wellness seekers in the community- based health services. The deep learning architecture is able to infer the possible diseases given the queries of health seekers. This approach also permits unsupervised feature learning from other wide range of syndrome kinds. The performance of the established system by means of increasing the number hidden layers is shown below.

The following table shows the performance augmentation when the number of hidden layers is increased. A structure containing three hidden layers shows ultimate performance of 98.21%.

Table 1: Disease Inference Performance with Various Numbers of Hidden Layers

Number of Layers	Performance on Dataset
Structure with One hidden layer	89.00%
Structure with two hidden layer	93.13%
Structure with three hidden layer	98.21%

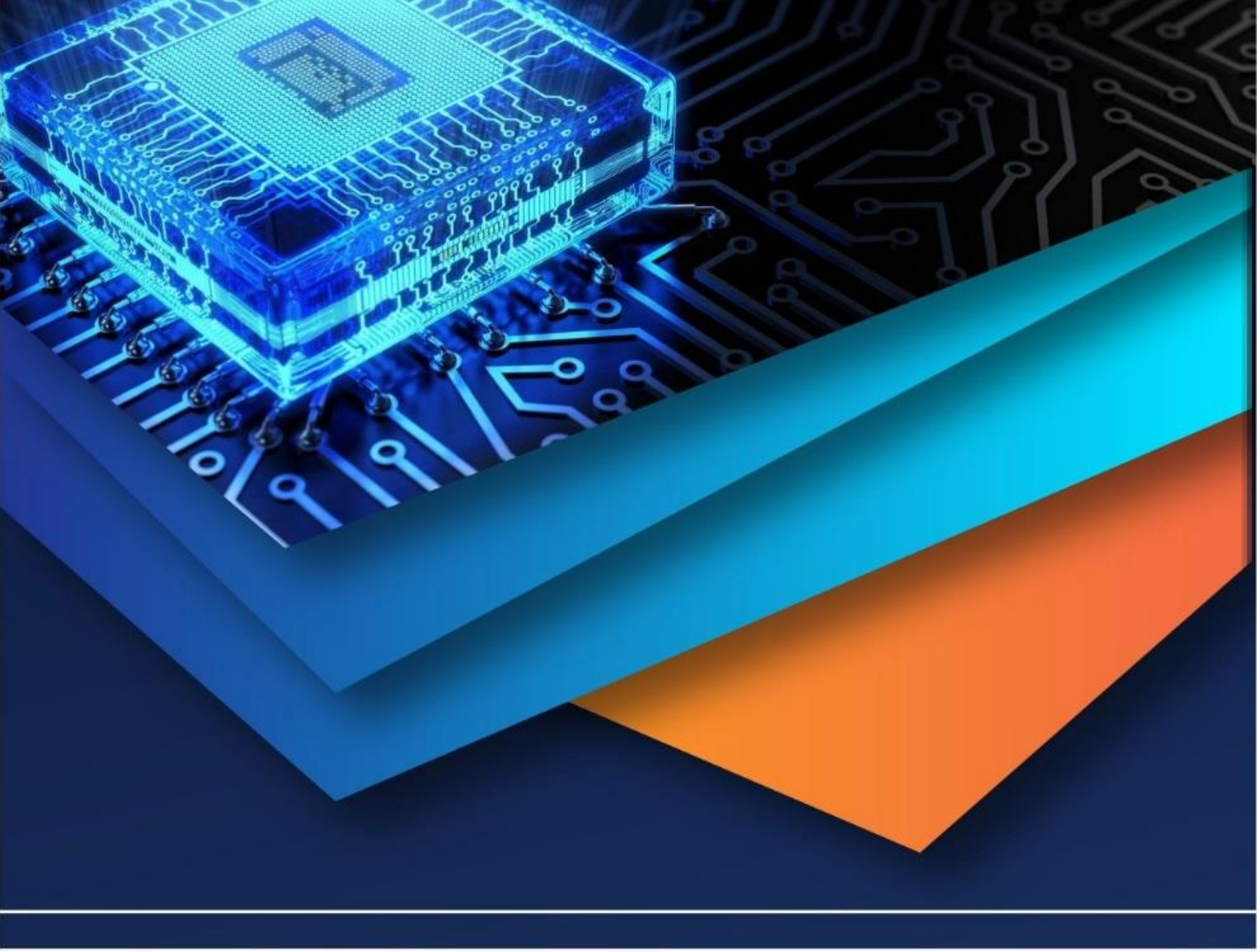
V. CONCLUSION

This paper conveys a study of several approaches used for classifying medical relationships and confines the inequalities between online health seeker and provider. The rewards and shortcomings of each procedure are calculated. The association between different tactics is quantified. It is stated that textual replies are felt as ideal ones by the online health seeker and the only limitation is the medical jargons gap among health seeker and doctor. This restriction is reduced in paper [8] which deals with numerous clinical text mining methods to augment answers in community based health care structures that shows a great performance development. The stretched out assessments on a real world dataset validate that scheme mentioned in [8] is able to produce promising performance when equated to the prevalent coding procedures. The widespread tactic is unsupervised and holds potential to deal with the large-scale data.

REFERENCES

[1] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the vocabulary gap between health seekers and healthcare knowledge," IEEE Trans. Knowl. Data Eng., vol. 27, no. 2, pp. 396–409, Jun. 2014
[2] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua, "Beyond text qa: Multimedia answer generation by harvesting web information," IEEE Trans. Multimedia, vol. 15, no. 2, pp. 426–441, Feb. 2013.
[3] Jimeno Yepes, Antonio; Berlanga, Rafae ,"Knowledge based word-concept model estimation and refinement for biomedical text mining," Journal of Biomedical Informatics (Elsevier) ,vol.53,pp. 300-307,2015.

[4] Ben Abacha, Asma; Zweigenbaum, Pierre "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," Journal of Information Processing & Management (Elsevier), vol.51,no. 5,pp- 570-594 ,2015.
[5] Frunza, Oana; Inkpen, Diana; Tran, Thomas," A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts," IEEE Transactions on Knowledge and Data Engineering; Tran, vol.23, no.6, pp. 801-814, 2011.
[6] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. Laine, "A framework for mining signatures from event sequences and its applications in healthcare data," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 2, pp. 272–285, Feb. 2013
[7] Fleuren, Wilco W.M.; Alkema, Wynand, "Application of text mining in the biomedical domain," Journal of Methods (Elsevier), vol.74, pp. 97-106, 2015.
[8] Liqiang Nie, Meng Wang, Luming Zhang, Shuicheng Yan, Member, Bo Zhang and Tat-Seng Chua , "Disease Inference from Health-Related Questions via Sparse Deep Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 8, 2107-2119, august 2015.
[9] Zhang, Shaodian; Elhadad, Noémie, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," Journal of Biomedical Informatics (Elsevier) vol. 46, no. 6, pp. 1088-1098, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY