DETECTION AND NORMALIZATION OF MEDICAL SEMANTICS FOR PERSONALIZED HEALTH SERVICES

Arunjunai Selvi.M
B.E scholar,
Computer Science and Engineering,
Velammal Institute of Technology,
Chennai ,India.

Deekshitha.L
B.E scholar,
Computer Science and Engineering,
Velammal Institute of Technology,
Chennai ,India.

Nithiyalakshmi.V
B.E scholar,
Computer Science and Engineering,
Velammal Institute of Technology,
Chennai ,India.

Tamilarasi.K
Assistant Professor,
Computer Science and Engineering,
Velammal Institute of Technology,
Chennai ,India.

Abstract: The process of analyzing data from different perspectives and summarizing it into useful information is defined as data mining (knowledge discovery).It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. We focused on bridging the vocabulary gap between health seekers and providers to code the medical records by utilizing local mining and global learning approaches. Generally, the community-based healthcare service including HealthTap2, HaoDF3 and WebMD4 are disseminating personalized health knowledge and connecting patients with doctors worldwide via question answering. It is difficult to avail online always by health service providers (Doctors). Further, most of the previous work simply utilizes the external medical dictionary to code the medical records rather than considering the corpus-aware terminologies which is focused on hospital generated health data or health provider released sources by utilizing either isolated or loosely coupled rule-based and machine learning approaches. We propose a novel scheme that is able to code the medical records with corpus-aware terminologies .Additional to these, we incorporated Separate Server Implementations for Local Mining and Global Learning, Medi Net Library, NLP Techniques, PDF Searching using Indexing and android applications for users.

Keywords : Corpus-aware terminologies, Medi Net Library, NLP Technique, Local Mining, Global Learning

I. INTRODUCTION

Recent studies have revealed that patients prefer gone online to figure out their medical conditions rather than embracing doctor's advice. The community based health care services are disseminating personalized health knowledge and connecting patients with doctors worldwide via question answering. These forums are very attractive to both professionals and health seekers. For professionals, they are able to increase their reputations among their colleagues and patients, strengthen their practical knowledge from interactions with other renowned doctors, as well as possibly attract more new patients.

In many cases, the community generated content, however, may not be directly usable due to the vocabulary gap. Users with diverse backgrounds do not necessarily share the same vocabulary. Take Health- Tap as an example, which is a question answering site for participants to ask and answer health-related questions.

The questions are written by patients in narrative language. The same question may be described in substantially different ways by two individual health seekers. On the other side, the answers provided by the well-trained experts may contain acronyms with multiple possible meanings, and no standardized terms. For example, "heart attack" and "myocardial disorder" are employed by different experts to refer to the same medical diagnosis. It was shown that the inconsistency of community generated health data greatly hindered data exchange, management and integrity. Even worse, it was reported that users had encountered big challenges in reusing the archived content due to the incompatibility between their search terms and those accumulated medical records. Therefore, automatically coding the medical records with standardized terminologies is highly desired.

II. RELATED WORK

Most of the current health providers organize and code the medical records manually [1]. A system named IndexFinder [2], proposed a new algorithm for generating all valid UMLS terminologies by permuting the set of words in the input text and then filtering out the irrelevant concepts via syntactic and semantic filtering.

A proposal in [3], instead of just converting the corpus data to terminologies, suggested users with appropriate medical terminologies for their personal queries. It integrated UMLS, WordNet as well as Noun Phraser to capture the semantic meaning of the queries. Suominen et al. [4] introduced a cascade of two classifiers to assign diagnostic terminologies to radiology reports. In their model, when the first classifier made a known error, the output of the second classifier was used instead to give the final prediction.

Beyond medical domain, several prior efforts of corpus alignment and gap bridging have been dedicated to other verticals. Chen et al. [5] derived an integrated model that jointly aligns bilingual named entities between Chinese and English news. Another example, the music semantic gap between textual query and audio content was remedied by annotation with concepts [6]. This approach can hardly be applied to medical terminology assignment directly due to the differences in modalities and content structures. Besides, it targets at labeling music entities with common noun and adjective phrases, while our approach focuses on terminologies only.

Similar to our scheme, Pakhomov et al. [7] attempted to improve the coding performance by combing the advantages of rule-based and machine learning approaches. It described Autocoder, an automatic encoding system implemented at Mayo clinic. Autocoder combines example based rules and a machine learning module using Na€ıve Bayes. However, this integration is loosely coupled and the learning model cannot incorporate heterogeneous cues, which is not a good choice for the community-based health services.

Information technologies are transforming the ways healthcare services are delivered from doctors to patients seeking online information that concerns their health.. The questions are written in narrative language by patients and that same can be answered by two individual experts differently and may contain acronyms with possible meanings. To bridge this gap, a novel scheme to code the medical records by utilizing together local mining and global approaches [8].Research oriented to promote interoperability between controlled vocabularies is an extensively recognized topic in biomedical community.

Mapping terminologies presents a serious difficulty in considerable volume of research in different fields. Two problems making interoperability difficulty, one is the informal treatment of relationships in biomedical terminologies. The second is the lack of automated methods simplifying the mapping process.

A proposal in [9] provides automated approach to mapping external terminologies to the Unified Medical Language(UMLS).A proposal in [10] provides supervised learning approaches to automatically assign international classification of diseases (ninth revision) - clinical modification (ICD-9-CM) codes to EMRs by experimenting with a large realistic EMR dataset.In work [11], they compared two efficient algorithms for diagnosis coding on a large patient dataset are Support Vector Machines(SVM) and Bayesian Ridge Regression.

III. EXISTING SYSTEM

Various types of Q and A blogs including General ,Technical and Medical are available where Medical related blogs are very less in number. Also the Questions posted in such blogs are likely to incur more time for being answered as the answers can only be posted by Medical Experts and Doctors who does not find time to spend online. General blogs and Technical blogs are very usual and any user can suggest answers for questions posted in them. Medical blogs are restricted to general user answers as it might lead to negative results on human lives.

It is worth mentioning that there already exist several efforts dedicated to research on automatically mapping medical records to terminologies. Most of these efforts, however, focused on hospital generated health data or health provider released sources by utilizing either isolated or loosely coupled rule-based and machine learning approaches. Compared to these kinds of data, the emerging community generated health data is more colloquial, in terms of inconsistency, complexity and ambiguity, which pose challenges for data access and analytics. Further, most of the previous work simply utilizes the external medical dictionary to code the medical records rather than considering the corpus-aware terminologies. Their reliance on the independent external knowledge may bring in inappropriate terminologies. Constructing a corpus aware terminology vocabulary to prune the irrelevant terminologies of specific dataset and narrow down the candidates is the tough issue we are facing. Major problem definition includes time delay for answers, ambiguity in key terms, vocabulary gap, Lack of artificial intelligence and lack of machine learning. In addition, the varieties of heterogeneous cues were often not adequately exploited simultaneously. Therefore, a robust integrated framework to draw the strengths from various resources and models is still expected.

IV. PROPOSED SYSTEM

We propose a novel scheme that is able to code the medical records with corpus-aware terminologies. The proposed scheme consists of two mutually reinforced components, namely, local mining and global learning. This is the work on automatically coding the community generated health data, which is more complex, inconsistent and ambiguous compared to the hospital generated health data. It proposes the concept entropy impurity approach to comparatively detect and normalize the medical concepts locally, which naturally construct a corpus-aware terminology vocabulary with the help of external knowledge. It builds a novel global learning model to collaboratively enhance the local coding results. This model seamlessly integrates various heterogeneous information cues. In addition to these we incorporated Separate Server Implementations for Local Mining and Global Learning and we introduced Medi Net library for medical term detection and Normalization, NLP techniques for processing narrative language which includes POS tagging, stemming process, error checking and PDF searching using Indexing approach for easy retrieval of PDF contents and using machine in responding by Artificial Intelligence.

A. LOCAL MINING

This section details the local mining approach. To accomplish this task, we establish a tri-stage framework. Specifically, given a medical record, we first extract the embedded noun phrases. We then identify the medical concepts from these noun phrases by measuring their specificity. Finally, we normalize the detected medical concepts to terminologies.

1. NOUN PHRASE EXTRACTION

To extract all the noun phrases, we initially assign part of speech tags to each word in the given medical record by Stanford POS tagger. We then pull out Noun phrases by WordNet. Noun phrase has a pattern which is formulated as:
(Adjective |Noun)*(Noun Preposition)? (Adjective |Noun)*Noun.

2. MEDICAL CONCEPT DETECTION AND NORMALIZATION

This stage aims to differentiate the medical concepts from other general noun phrases. Inspired by the efforts in ,we assume that concepts that are relevant to medical domain occur frequently in medical domain and rarely in non-medical ones. And also Medical concepts are defined as medical domain specific noun phrases; we cannot ensure that they are standardized terminologies. Take "birth control" as an example, it is recognized as a medical concept by our approach, but it is not an authenticated terminology. Instead, we should map it into "contraception". Therefore, it is essential to normalize the detected medical concepts according to the external suitable standardized dictionary and this normalization is the key to bridging the vocabulary gap. We incorporated Medi Net library which consists of medical terms with meanings for medical term detection and normalization.

B. GLOBAL LEARNING

The target of this section is to learn appropriate terminologies from the global terminology space T to annotate each medical record q in Q. Among existing machine learning methods, graph-based learning achieves promising performance .In this work, we also explore the graph-based learning model to accomplish our terminology selection task, and expect this model is able to simultaneously consider various heterogeneous cues, including the medical record content analysis, terminology-sharing networks, and the inter-expert as well as inter-terminology relationships. We will first introduce relationship identification and then we detail how to use our proposed model to link the underlying connected medical records. Next, we present the optimal solution for our learning model followed by the label bias estimation. Finally, we discuss the scalability of our method.

1. RELATIONSHIP IDENTIFICATION

The inter-terminology and inter-expert relationships are not intuitively seen or implied from medical records. We thus call them as implicit relationships. This subsection aims to introduce how to discover these kinds of relationships.

2. INTER-TERMINOLOGY RELATIONSHIP

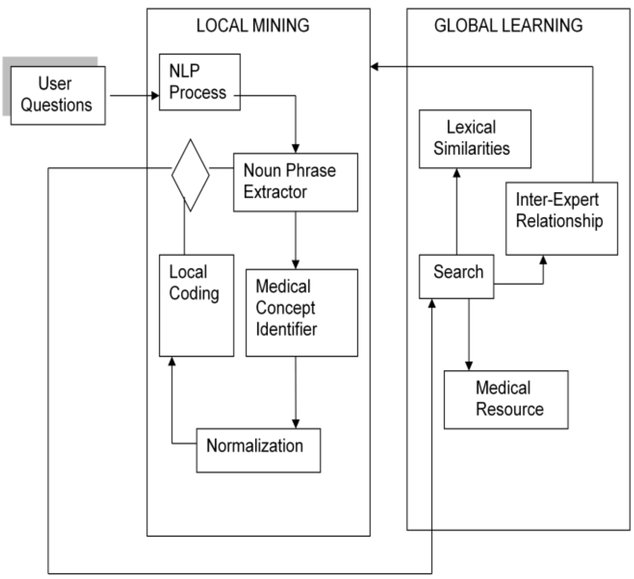
The medical terminologies in SNOMED CT are organized into acyclic taxonomic (is-a) hierarchies. For example, "viral pneumonia" is-an "infectious pneumonia" is-"pneumonia" is-a "lung disease". Terminologies may also have multiple parents. For example, "infectious pneumonia" is also a child of "infectious disease".
The medical terminology hierarchy will enhance our scheme in two ways. First, it tackles the granularity mismatch problem, where the terminologies found in the medical records are very detailed and specific, while those in the query may be more general and high-level. This is achieved by rewarding the ancestral nodes with appropriate weights. Second, the hierarchical relationships boost the coding accuracy via filtering out the sibling terminologies.

3. INTER-EXPERT RELATIONSHIP

The inter-expert relationships will be viewed stronger if the experts are professionals in the same or related specific medical areas. This is reflected by their historical data, i.e., the number of questions they have co-answered.

V. ARCHITECTURE

The proposed scheme consists of two mutually reinforced components, namely, local mining and global learning. Local mining aims to locally code the medical records by extracting the medical concepts from individual record and then mapping them to terminologies based on the external authenticated vocabularies. We establish a tri-stage framework to accomplish this task, which includes noun phrase extraction, medical concept detection and medical concept normalization. As a by-product, a corpus-aware terminology vocabulary is naturally constructed, which can be used as terminology further learning in the second component. However, local mining approach may suffer from the problem of information loss and low precision due to the possible lack of some key medical concepts in the medical records and the presence of some irrelevant medical concepts. We thus propose global learning to complement the local medical coding in a graph-based approach. It collaboratively learns missing key concepts and propagates precise terminologies among underlying connected records over a large collection.



VI.EXPERIMENT

We did our experiment in four modules which includes Q and A Application, NLP Process, Bridging Gap, and Machine Learning. Here the modules are explained as follows:

A. Q AND A APPLICATION.

Generally, In Existing Web Applications the Questions posted by the users are answered by the Other User which might result in redundancy and user unbelivability especially for medical related doubts, clarifications and questions. So a medical Experts who can give believable answers should be available all the time which is practically not possible and time Consuming .So we build a Efficient Q and A Scheme which could give Instant Answers Analyzing the Users Objective behind the Question.

B. NLP PROCESS

The User posts Questions for instant answers is processed by a natural language processing technique so that the proper meaning would be revealed. The Nlp Process comprises a several steps. Of which Parts Of Speech Tagging (POST) results in Phrases and Nouns Extraction. The Keywords thus Extracted is subject to Stemming Process which eliminates the Stop words in the sentence and also trims the keyword for Base Word.

C. BRIDGING GAP

The Proper meanings will be analyzed with a English Dictionary and the Medical Terms will be Normalized based on Domain Specific Knowledge. Medical Terminologies were Collected and grouped so that the checking with the synonyms of keywords could result in Normalization. The Normalized words will be checked for Contradictions with medical terminologies and the related answers will be queried from Local Mining Database.

D. MACHINE LEARNING

Machine Learning in Our Approach is achieved by the use of Local Mining and Global Learning techniques. Local Mining database gets updated by the Global learning data's once user posts a newer Kind of Query to the Answering System. The Global learning Comprises a large collection of Medical Related Resources in its backend which helps to retrieve a related resource to the Query based on terminology keywords. This Search is completely indexed an 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. Once Experts reviewed the query the answers not only dispatches to the Medical Seekers and also updates the Local Mining Database for future instant retrieval to the related Query from other Users.

VII.ALGORITHMS

A. CLUSTERING

We use clustering algorithms to group the similar medical terms under a common domain. For example allergy, skin, egg, rashes etc.. are grouped under Allergy. The objective to do this is we can find out the domain under which the posted question comes under.

B. NLP

We have attempted to automatically convert free medical texts into medical terminologies / ontologies by combining several natural language processing methods, such as Noun phrase extraction, stemming, Spelling and grammar check, morphological analysis, lexicon augmentation, term composition and negation detection.

C. INDEXING

We have used Lucene Indexing for fast and efficient PDF searching and downloading. Any PDF of any disease can be retrieved by the user which consists of all information about the disease, cause and how to control it. All the PDF sources are converted into indexed files which is machine readable, so that the PDF files can be searched easily.

D. BUBBLE SORT

We uses Bubble sort for prioritizing process, so that the questions which are frequently searched are come in the top list in database so the answers can be responded very efficiently and with small cost.

VIII. CONCLUSION

This paper presents a medical terminology assignment scheme to bridge the vocabulary gap between health seekers and healthcare knowledge. The scheme comprises of two components, local mining and global learning. The former establishes a tri-stage framework to locally code each medical record. However, the local mining approach may suffer from information loss and low precision, which are caused by the absence of key medical concepts and the presence of the irrelevant medical concepts. This motivates us to propose a global learning approach to compensate for the insufficiency of local coding approach. The second component collaboratively learns and propagates terminologies among underlying connected medical records. It enables the integration of heterogeneous information. Extensive evaluations on a real-world dataset demonstrate that our scheme is able to produce promising performance as compared to the prevailing coding methods. More importantly, the whole process of our approach is unsupervised and holds potential to handle large-scale data. Local Mining Gives direct Answers and Global Learning is implemented as a Search Engine. Machine Learning improves system performance.

IX. REFERENCES

[1] AHIMA e-HIM Work Group on Computer-Assisted Coding, "Delving into computer-assisted coding," J. AHIMA, vol. 75,pp. 48A–48H, 2004.

[2] Q. Zhou, W. W. Chu, C. Morioka, G. H. Leazer, and H. Kangarloo,"Indexfinder: A method of extracting key concepts from clinical texts for indexing," in Proc. AMIA Annu. Symp., 2003, pp. 763–767.

[3] G. Leroy and H. Chen, "Meeting medical terminology needs-the ontology-enhanced medical concept mapper," IEEE Trans. Inf.Technol. Biomed., vol. 5, no. 4, pp. 261–270, Dec. 2001.

[4] H. Suominen, F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S.Salanter, and T. Salakoski, "Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: A method description," in Proc. ICML Workshop Mach. Learn. Health-Care Appl., 2008.

[5] Y. Chen, Z. Chenqing, and K.-Y. Su, "A joint model to identify and align bilingual named entities," Comput. Linguistics, vol. 39, no. 2,pp. 229–266, 2013.

[6] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," IEEE Trans. Multimedia, vol. 13, no. 2, pp. 303–319, Apr. 2011.

[7] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," J. Amer. Med. Inf.
Assoc., vol. 13, no. 5, pp. 516–525, 2006.

[8] Liqlang Nie,Yi-Liang Zhao,Mohammad Akbari,Jialie Shen and Tat-Seng Chua, "Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge",IEEE Trans. On Kno. And DataEngg.,vol 27,no.2,Feb.2015

[9] Maria Taboada,Rosario Lalin and Diego Martinez, "An Automated Approach to Mapping External Terminologies to the UMLS",IEEE Trans. Biomed. Engg.,vol 56,no.6,June 2009.

[10] Ramakanth Kavuluru,Anthony Rios and Yuan Lu, "An Empirical Evaluation of Supervised Learning Approaches in Assigning Diagnosis Codes to Electronic Medical Records".

[11] Lucian Vlad Lita,Shipeng Yu and Stefan Niculescu, "Large Scale Diagnostic Code Classification for Medical Patient Records".