Health Seekers Understanding Medical Terminology from the Health Care Experts

R.Nithya[#1] and I.Anette Regina[#2]

[1]*M.Phil Scholar, Department of Computer Science, Muthurangam Govt Arts College,*

[2,] *Associate Professor, Department of Computer Science, Muthurangam Govt Arts College,*

Vellore, Tamilnadu, India.

[1]*nirmala.nithi@gmail.com*

[2]*anette1967cs@gmail.com*

***Abstract*— Recent year's health service websites are dramatically developed. In such health services Medical term and technical verbal's play vital role between the health seekers and health care expert's knowledge has caught up data access. To viaduct this gap, this approach presents a scheme to label question answer (QA) pairs by together utilizing local mining and global learning approaches. Local mining attempts to the persons query can be processed and replayed with healthcare experts that can be no of experts. In local mining map all the user queries and no more same kind of reply from experts have it in local mining. However, it may suffer from data loss and inferior precision, which are caused by the nonappearance of key medical concepts and occurrence of unrelated medical terms. Information technologies are transforming the ways healthcare services are delivered, from patient's passively embracing their doctor's orders to patient's actively seeking online information that concerns their health. Some time that medical terminology not able to understandable normal health seekers. This research presents the new approach that provides the health seekers may convert medical terminologies into their local languages and they can easily understand the Information.**

***Keywords*— Home mining, Universal Learning.**

I. INTRODUCTION

Information technologies are transmuting the ways healthcare services are transported, from patients' submissively espousal their doctors' orders to patients' vigorously seeking online information that concerns their health. To better provide to health seekers, a increasing number of community-based healthcare services have twisted up, including HealthTap,2 HaoDF3 and WebMD.4 They are disseminating personalized health knowledge and connecting patients with health experts worldwide via question answering. [1], [2]. These forums are very gorgeous to both experts and health seekers. For experts, they are able to increase their statuses among their generations and patients, reinforce their applied information from communications with other well-known doctors, as well as perchance attract more new patients. For patients, these systems deliver nearly immediate and important responses particularly for composite and urbane problems.

Over times, a marvelous number of medical records have been accrued in their sources, and in most situations, users may straight trace good answers by searching from these record archives, rather than waiting for the experts' replies or looking through a list of possibly appropriate documents from the Web.

In many cases, the community produced content, however, may not be straight usable due to the vocabulary gap.

Users with diverse circumstances do not essentially share the same vocabulary. Take Health Tap as an example, which is an enquiry answering site for members to ask and answer health-related questions. The questions are written by patients in description language. The same question may be described in significantly diverse ways by two distinct health seekers. On the other side, the replies provided by the well-trained specialists may contain acronyms with multiple thinkable meanings, and non-standardized terms.

Recently, some sites have encouraged experts to interpret the medical archives with medical concepts. However, the tags used often vary wildly and medical concepts may not be remedial terminologies [3]. For instance, "heart attack" and "myocardial disorder" are employed by unrelated experts to refer to the same medical analysis. It was shown that the discrepancy of public produced health data greatly delayed data exchange, organization and integrity [4]. Even worse, it was described that users had encountered big trials in reusing the archived content due to the mismatch between their search terms and those accumulated medical records [5]. Therefore, mechanically coding the medical records with dependable terminologies is highly desired. It leads to a dependable interoperable way of indexing, stowage and combining across subjects and sites. In addition, it facilitates the medical record retrieval via connecting the vocabulary gap between enquiries and records.

It is worth mentioning that there already exist several efforts dedicated to research on automatically mapping medical records to terminologies [6], [7], [8], [9], [10], [11].

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. 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.

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, Home mining and Universal learning. Home mining aims to Homely 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 space for further learning in the second component.

However, Home 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 Universal learning to complement the Home 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. Besides the semantic similarity among medical records and terminology-sharing network, the inter-terminology and inter-expert relationships are seamlessly integrated in the proposed model. The inter-terminology relationships are mined by exploiting the external well-structured ontology, which are able to alleviate the granularity mismatch problems and reduce the irrelevant sibling terminologies. The inter expert relationships are inferred from the experts' historical data. It may be capable of excluding a wealth of domain-specific context information. Specifically, the medical professionals who are frequently responding to the same kinds of questions probably share highly overlapping expertise, and thus the questions they answered can be regarded as semantically similar to a certain extent. Extensive evaluations on the real-world dataset demonstrate that our proposed scheme can achieve significant gains in medical terminology assignment. Meanwhile, the whole process of our proposed approach is unsupervised and it holds potential to handle large-scale data.

The main contributions of this work are threefold:

- To the best of our knowledge, this is the first 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 (CEI) approach to comparatively detect and normalize the medical concepts Homely, which naturally construct a
- Corpus-aware terminology vocabulary with the help of external knowledge.

II. RELATED WORKS

Most of the current health providers organize and code the medical records manually [3]. This workflow is extremely expensive because only well-trained experts are properly competent for the task. Therefore, there is a growing interest to develop automated approaches for medical terminology assignment. The existing techniques can be categorized into two categories: rule-based and machine learning approaches.

Rule-based approaches play a principle role in medical terminology assignments [6], [7], [8]. They generally discover and construct effective rules by making strong uses of the morphological, syntactic, semantic and pragmatic aspects of natural language. It has been found that these methods have significant positive effects on the real systems [12]. Back in 1995, Hersh and David [13] designed and developed a system, named SAPPHIRE, which automatically assigned UMLS5 terminologies to medical documents using a simple lexical approach. Around one decade later, a system named Index Finder [14], 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. Most recently, several efforts [12], [15], [16], [17] have attempted to automatically convert free medical texts into medical terminologies/ontologies by combining several natural language processing methods, such as stemming, morphological analysis, lexicon augmentation, term composition and negation detection.

However, these methods are purely applicable to well-constructed discourses. A proposal in [4], 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. However, an implicit assumption of this work is that the sources to be searched must be well presented using a standardized medical vocabulary. Obviously, this is not applicable to the community generated medical sources. In summary, even though rule-based methods are fast and suitable for real-time applications, the rule construction is challenging and the performance varies from different corpus.

Machine learning approaches build inference models from medical data with known annotations and then apply the trained models to unseen data for terminology prediction [6], [18]. The research can be traced back to the 1990 s, where Larkey and Croft [10] have trained three statistical classifiers and combined their results to obtain a better classification in 1995. In the same year, support vector machine (SVM) and Bayesian ridge regression were first valuated on large-scale dataset and obtained promising performance [9]. Following that, a hierarchical model was studied in [19], which exploited the structure of ICD-9 code set and demonstrated that their approach outperformed the algorithms based on the classic vector space model. About ten years later, Suominen et al. [11] 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. Yan et al. [20] proposed a multi-label large-margin formulation that explicitly incorporated the inter-terminology structure and prior domain knowledge simultaneously. This approach is feasible for small terminology set but is questionable in real-life settings where thousands of terminologies need to be considered.

Similar to our scheme, Pakhomov et al. [21] 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 Naıve 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.

Beyond medical domain, several prior efforts of corpus alignment and gap bridging have been dedicated to other verticals. Chen et al. [22] derived an integrated model that jointly aligns bilingual named entities between Chinese and English news. The work in [23] bridged the management research-practice gap by describing their experiences with the network for business sustainability. A game platform was designed in [24] and was demonstrated how to enhance the inter-generation cultural communication in a family. These diverse efforts are all heuristic. Their rules and patterns are domain specific and cannot be generalized to other areas. Another example, the music semantic gap between textual query and audio content was remedied by annotation
with concepts [25]. 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.

III. HOME MINING

Medical concepts are defined as medical domain-specific noun phrases, and medical terminologies are referred to as authenticated phrases by well-known organizations that are used to accurately describe the human body and associated components, conditions and processes in a science- based manner. This section details the home 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.

The entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes.

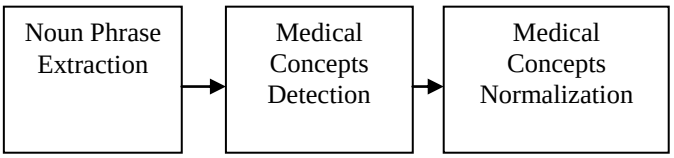

Fig 1. Home Mining steps.

A. *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.6 we then pull out sequences that match a fixed pattern as noun phrases. This pattern is formulated as follows:

(Adjective | Noun)(Noun Preposition)*
*? (Adjective | Noun)*Noun*

The above regular expression can be intuitively interpreted as follows. The noun phrases should contain zero or more adjectives or nouns, followed by an optional group of a noun and a preposition, followed again by zero or more adjectives or nouns, followed by a single noun. A sequence of tags matching this pattern ensures that the corresponding words make up a noun phrase. For example, the following complex sequence can be extracted as a noun phrase: "ineffective treatment of terminal lung cancer". In addition to simply pulling out the phrases, we also do some simple post processing to link the variants together, such as singularizing plural variants.

B. *Medical Concept Detection*

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.

C. *Medical Concept Normalization*

In this work, we use SNOMED CT because it provides the core general terminologies for the electronic health record and formal logic-based hierarchical structure. Home mining terminologies may suffer from various problems. The first problem is incompleteness. This is because some key medical concepts not explicitly present in the medical records are excluded. The second one is the lower precision. This is due to some irrelevant medical concepts explicitly embedded in the medical records, and is mistakenly detected and normalized by the Home approach. Another issue, which deserves further discussion here, is the terminology space. It may result in the deterioration in coding performance in terms of efficiency and Effectiveness.

IV. GRAPH BASED UNIVERSAL LEARNING

Let P= {p_1, p_2... p_N} and T= {t_1, t2… t_M } respectively denotes a repository of medical records and their associated homely mined terminologies. The target of this section is to learn appropriate terminologies from the Universal terminology space T to annotate each medical record p in P. Among existing machine learning methods, graph-based learning achieves promising performance [28], [29].

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.

A. *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. The search results are viewed into multiple languages.

V.EXPERIMENTS

In this experimental result, first demonstrate that our algorithm using .NET to reduce the information gap between domains during the process of including the unlabeled data from the related domains. Randomly selects sample for these tasks and for each of the data sets display the performance, together with their corresponding margin sizes. For each iteration, include the top level unlabeled data that are closest to the decision boundary for information gap.

This approach gives balance between information sharing between health seekers health care experts. Some approach gives only English dictionary conversion from the data set but it's not very effective whether all health seekers not well known in English pronouncing person while this approach has given dataset to some of multiple local languages that can be very easy to under stable manner to the health seekers.

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

.

The graph represents evaluation results. When compare to local mining global learning separately, the combine approach of the local mining and global learning is producing good results within the performance parameters like Precision, Recall and Accuracy.

Fig.2. Getting Search Results in Local Languages.

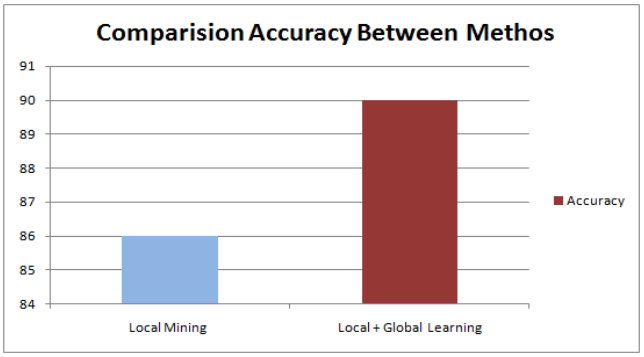

Fig.3. Precision Performance between Approaches.

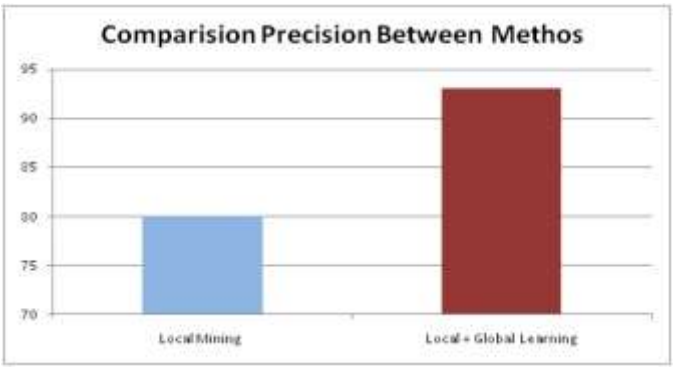

Fig 4. Accuracy comparison.

V. CONCLUSION

The proposed approach is primarily bearing in mind over the concept of shared approach of local mining and global learning. This shared approach is serving the patients to get benefit by the expert's implication, the patients are also not able to get the health care experts always and health care experts also haven't all the records of any particular patient, every health care expert have more number of patients. The proposed approach consist of a combined approach within the local mining and global learning, where the corpus aware terminology is being used for making a communication between the medical support seeker and the medical care

providers. The corpus terminology is having the combined approaches of local mining and global learning, where the approach of local mining undergoes within the process of stemming, noun phrase extraction, spell check, normalization and detection of medical concept. The global learning maps the query against the indexed text or keyword that is pertinent to the medical records. The query is being mapped within the local database and health seekers. The output is being formed based on the patients query and construct no of local languages. In the future, search can be conceded out on how to flexibly organize the unstructured medical content into user needs-aware ontology by leveraging the recommended medical terminologies.

REFERENCES

[1]. Nie L, Zhao Y-L, Akbari M, Shen J, Chua T-S." Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge". IEEE Transactions on Knowledge and Data Engineering. 2014; 27(2):396-409.

[2]. Alpay L, Verhoef J, Xie B, Te'eni D, Zwetsloot-Schonk JH. "Current Challenge in Consumer Health Informatics: Bridging the Gap between Access to Information and Information" Understanding Biomed Inform Insights in PMC. 2010; p. 1-10.

[3]. Nie L, Wang M, Zhang L, Yan S. "Disease Inference from Health-Related Questions via Sparse Deep Learning". IEEE Transactions on Knowledge and Data Engineering. 2015; 27(8):2107-119.

[4]. Nie L, Akbari M, Li T, Chua T-S. "A joint local-global approach for medical terminology assignment". Proc. Int. Conf. ACM SIGIR. 2014; p. 24-26.

[5]. Kim MY, Goebel R. Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking. 2010 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB), Corfu. 2010; p. 1-5.

[6]. Cilibrasi RL, Vitanyi PMB. The google similarity distance. IEEE Transactions on Knowledge and Data Engineering. 2007; 19(3):370-83.

[7]. Dozier C, Kondadadi R, Al-Kofahi K, Chaudhary M, Guo X. "Fast Tagging of Medical Terms in Legal Text". In Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL 2007). 2007; p. 253-60.

[8]. Lita LV, Yu S, Niculescu S, Bi J." Large scale diagnostic code classification for medical patient records". In Proc. Conf. Artif. Intell. Med. 1995; p. 877-82.

[9]. Terol RM, Martinez-Barco P, Palomar M. "A knowledge based method for the medical question answering problem". Computers in Biology and Medicine. 2007; 37(10):1511-21.

[10]. Niu Y, Hirst G. "Analysis of semantic classes in medical text for question answering". Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains. 2004; p. 1-8.

[11] 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.

[12] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt, "Exploiting medical hierarchies for concept-based information retrieval," in Proc. Australasian Document Comput. Symp., 2012,pp. 111–114.

[13] E. J. M. Laurıa and A. D. March, "Combining Bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis," J. Data Inf. Quart., vol. 2, no. 3, p. 13, 2011.

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

[15] Y. Wang and J. Patrick, "Mapping clinical notes to medical terminology at point of care," in Proc. Workshop Current Trends Biomed. Natural Lang. Process., 2008, pp. 102–103.

[16] S. Hina, E. Atwell, and O. Johnson, "Semantic tagging of medical narratives with top level concepts from SNOMED CT healthcare data standard," Int. J. Intell. Comput. Res., vol. 2, pp. 204–210, 2010.

[17] H. Stenzhorn, E. Pacheco, P. Nohama, and S. Schulz, "Automatic mapping of clinical documentation to SNOMED CT," Studies Health Technol. Inform., vol. 158, pp. 228–232, 2009.

[18] K. Crammer, M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll, "Automatic code assignment to medical text," in Proc. Workshop Biol., Translational, Clinical Lang. Process., 2007, pp. 129–136.

[19] L. R. S. de Lima, A. H. F. Laender, and B. A. Ribeiro-Neto, "A hierarchical approach to the automatic categorization of medical documents," in Proc. Int. Conf. Inf. Knowl. Manag., 1998, pp. 132–139.

[20] Y. Yan, G. Fung, J. G. Dy, and R. Rosales, "Medical coding classification by leveraging inter-code relationships," in Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2012, pp. 193–202.

[21] 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.

[22] 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.

[23] P. Bansal, S. Bertels, T. Ewart, P. MacConnachie, and O. James, "Bridging the research–practice gap," Acad. Manag. Perspectives, vol. 26, pp. 73–91, 2012.

[24] N. Chu, Y. Choi, J. Wei, and A. Cheok, "Games bridging cultural communications," in Proc. IEEE Universal Conf. Consumer Electron., 2012, pp. 329–332.

[25] 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.

