Jun. 30

International Journal of advanced studies in Computer Science and Engineering
IJASCSE, Volume 4, Issue 6, 2015

Local-Global Learning Approaches for Bridging Vocabulary Gap in Health Care

Ms. Ashwini Rewatkar
Department of Computer Engineering
D.Y. Patil College of Engineering, Akurdi
Savitribai Phule Pune University, Pune, India

Mrs. M. A. Potey
Department of Computer Engineering
D.Y. Patil College of Engineering, Akurdi
Savitribai Phule Pune University, Pune, India

***Abstract*— Vocabulary gap between health seekers and community generated knowledge in community-based health services has hampered data access. This paper presents a scheme to bridge this gap using local mining and global learning approaches. Local mining extracts medical concepts from the Question Answering pair itself and then maps to authenticated terminologies. Global learning inclined towards enhancing the local mining via collectively discovering missing key terminologies and keeping off the irrelevant terminologies by analyzing the social neighbours. The upswing of digital technologies has transformed the doctor-patient relationships. In this era of web, when people conflict with their health concerns then generally most of them explore the Internet to research the problem before and after they see their doctors. A dictionary of vocabulary is build up as a by-product and which is used as the terminology space for global learning.**

Keywords- Question-Answering (QA), Local-Global Learning Approach, Health Care, Natural Language Processing (NLP).

I. INTRODUCTION

Information technology is convenient for transforming the health care information from patients to doctors via question answering. Question answering (QA) is the approach for automatically answering a question posed in natural language. The medical QA system is able to answer medical questions according to universal question taxonomy. Compared to keyword-based search systems, it highly facilitates the communication between human and computer by generally stating users intention in lucid sentences. QA's performance is hampered by complicated natural language processing (NLP) techniques. The system take input as, a natural language question and return a precise words that provide the answers.

Question answering forums attract the consideration of both patients and doctors. The patients are provided with an quickly and trusted answer for simple as well as complex health concerns. The doctors are able to rise their reputation amongst their colleagues and patient, strengthen their practical knowledge from interactions with other prominent doctors, as well as possibly attract more new patients. There is a vocabulary gap between health seekers and health provider, to link this gap local mining and global learning approaches are used.

A tremendous number of medical records have been acquired in their repositories, and in most circumstances, user may precisely locate good answers by exploring from these record archives, rather than waiting for the experts browsing through an inventory of potentially relevant documents from the web. They associate user with provider. Few forthcoming community-based healthcare services are HealthTap, HaoDF and WebMD.

Generally, the community generated content, may not be directly accessible due to the vocabulary gap. Users do not share similar vocabulary. For e.g. HealthTap, is a question answering site for users to ask health related questions. The questions are written by our own words. The similar question may be described in substantially multiple ways by two individual health seekers. The answer provided by the well experts may contain phrase with different possible meanings, 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. They were reported that users had encountered big challenges in reusing the archived fulfilled due to the incompatibility between their search terms and those accumulated medical records. Automatically coding of medical records is immensely desired using standardized terminologies. It facilitates the medical record retrieval via linking the vocabulary gap between queries and archives. Several efforts earlier exist on automatically mapping medical records to terminologies then they focus on hospital generated health data. The community develop heath records is more general, in terms of inconsistency, complexity and ambiguity. Previously, they simply handled the external medical dictionary to code the medical records rather than considering the health related terminologies.

A. Natural Language Processing

Natural Language Processing is a theoretically inspired range of computational techniques for analyzing and representing generally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human like language processing for a range of applications. The goal of NLP as declared above is to accomplish human like language processing. The choice of the word, processing is very deliberate, and should not be recovered with

understanding. The field of NLP was basically referred to as Natural Language Understanding (NLU) of AI, it is well agreed today that while the goal of NLP is true Natural Language Understanding, that goal has not yet been achieved.

Naturally occurring texts can be of any language, mode, etc. Only requirement is that they be in a language used by humans to communicate with each other. Also, the word being analyzed should not be specifically constructed for the purpose of the analysis, but rather that the word is gathered from actual usage. It runs into many stages, especially tokenization, lexical, syntactic, semantic, and pragmatic analysis. Syntactic analysis provides an order and structure of each sentence in the word. Semantic analysis is to find the similar meaning, and pragmatic analysis is to determine the meaning of the word in context.

II. Related Work

Yiliang Zhao et.al.,[1] have classified approach into the following two categories. Health provider released sources by utilizing either isolated or loosely connects rule-based and machine learning approaches. Most of the ongoing health providers organize and code the medical records manually. This workflow is extremely expensive because only well trained experts are properly capable for the task. Therefore, there is a growing interest to build up automated approaches for medical terminology assignment. The existing techniques can be considered into two categories: rule-based and machine learning methods.

A. Rule Based Approaches

Rule-based approaches play a standard role in medical terminology assignments. They normally discover and construct effective rules by building strong uses of the morphological, syntactic, semantic and pragmatic aspects of natural language. It has been found that these approaches have significant positive effects on the real systems. Several efforts have attempted to automatically translate free medical texts into medical terminologies/ontologies by combining several natural language processing methods, such as stemming, morphological analysis, lexicon augmentation, phrase composition and negation detection.

Rule based approaches are fast and suitable for real time applications, the rule creation is challenging and the performance varies from different corpus. Back in 1995 Hersh et.al.[8] designed and developed a system, called SAPPHIRE, which automatically assigned UMLS5 terminologies to medical documents using a simple lexical method. Around one decade later, Zau et.al.[9] have proposed a new algorithm for developing all valid UMLS terminologies by permuting the set of words in the input text and then separating out the irrelevant concepts via syntactic and semantic filtering, the system called Index Finder. Most recently, several efforts [12] [13] [20] have attempted to automatically translate free medical texts into medical terminologies by associating several natural language processing methods, such as stemming, morphological analysis, lexicon augmentation, term composition and negation detection. However, these approaches are exactly applicable to well-constructed discourses. A proposal in [10] suggests, instead of just converting the corpus data to terminologies, recommended users with appropriate medical terminologies for their personal queries. It combined UMLS, WordNet as well as Noun Phrase to capture the semantic meaning of the queries.

B. Machine Learning Approaches

Machine learning methods build inference models from medical data with well-known annotations and then apply the trained models to unseen data for terminology prediction. The research can be traced back to the 1990s, where Larkey and Croft [6] have trained three statistical classifiers and integrated their results to obtain a better classification in 1995. In that year, support vector machine (SVM) and Bayesian ridge regression were first evaluated on large-scale dataset and retrieved promising performance. The structure of ICD-9 code set and determined that their method outperformed the algorithms based on the classic vector space model. Around ten years later, Suominen et.al.[18] introduced a cascade of two classifiers to assign diagnostic terminologies to radiology reports. In that model, when the first classifier made a well known error, the output of the second classifier was used instead to give the final prediction. Yan et.al.[2] proposed a multi-label large-margin formulation that explicitly integrated the inter-terminology structure and prior domain knowledge simultaneously. This method is feasible for limited terminology set but is questionable in real-life settings where thousands of terminologies need to be considered.

Pakhomov et.al.[4] attempted to increase the coding performance by combing the advantages of rule-based and machine learning approaches. It represents Autocoder, an automatic encoding system implemented at Mayo clinic. Autocoder integrates example-based rules and a machine learning module using Nave Bayes. However, this combination is loosely coupled and the learning model cannot integrate heterogeneous suggestion, which is not a valuable choice for the community based health services. Beyond medical domain, several prior efforts of corpus alignment and gap linking have been dedicated to other verticals, and return a precise text that provides the answers.

III. Method

To prevail over these limitations, we propose a novel scheme that is able to code the QA pairs with corpus-aware terminologies. As illustrated in Figure 1, the proposed method consists of two jointly augmented components that is, local mining and global learning.

A. Local Mining Approach

Medical concepts are referred to medical domain specific noun phrases and Medical terminologies are allude 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.

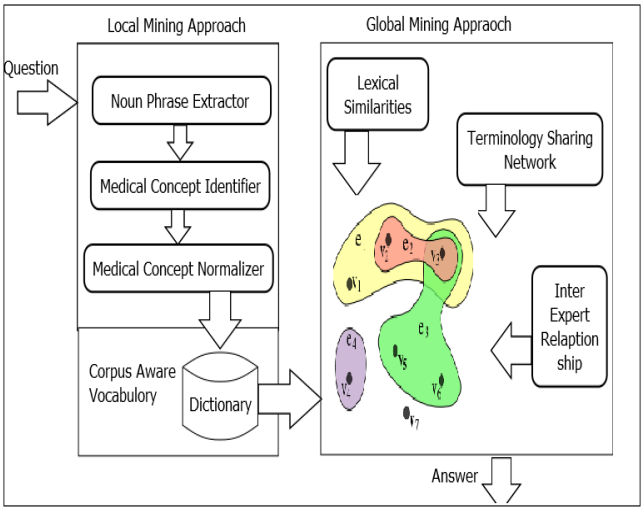

Fig. 1: System Architecture

It provides a tri-stage framework. Especially, given a medical record, it first extracts the embedded noun phrases, and then identifies the medical concepts from these noun phrases by measuring their specificity. Finally, it normalizes the detected medical concepts to terminologies.

1) Noun Phrase Extraction:

To extract entire the noun phrases, we originally assign part-of-speech tags to each word in the given medical record by Stanford POS tagger. And then pull out sequences that match a fixed pattern as noun phrases.

The regular expression can be possibly interpreted as follows. The noun phrases should consist of 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 texts make up a noun phrase. For e.g., the following complicated sequence can be extracted as a noun phrase: "ineffective treatment of terminal lung cancer". In addition to simply pulling out the phrases, and we also do some simple post processing to link the variants together, such as singularizing plural variants.

2) Medical Concept Detection:

In this stage aims to differentiate the medical concepts from other general noun phrases. Inspired by the efforts, it consider that concepts are relevant to medical domain occur frequently in medical domain and seldom in non-medical ones. The concept entropy impurity (CEI) is a similarity measure for domain relevance of a concept. In that impurity used c as concept, D1 and D2 represents our medical corpus and a general-domain corpus. P(Di j c) denotes the probability that a concept is related to a specified domain Di.

3) Medical Concept Normalization:

Although medical concepts are explain as medical domain-specific noun phrases, we cannot provide that they are standardized terminologies. For e.g., "birth control", that is recognized as a medical concept by local mining approach, but it is not an authenticated terminology. Rather than, it maps into "contraception". Hence, it is necessary to normalize the detected medical concepts according to the external appropriate standardized dictionary and this normalization is the key to linking the vocabulary gap.

Right now, there exist several authenticated vocabularies, including ICD7, UMLS, WordNet [21] and SNOMED CT [22]. These medical and clinical terminologies were invented in different times by different associations for multiple purposes. For e.g., ICD, in general it is used for external reporting requirements. In this work, we use WordNet because it provides the general terminologies for the electronic health record and formal logic-based hierarchical structure.

B. Global Learning Approach

Global learning is an important method, including local approach, and attempted to map the QA pairs directly to the entries in external dictionaries without any pruning. This method generally presents problems since the external dictionaries naturally cover relatively comprehensive terminologies and are far beyond the vocabulary scale of the given corpus. It may result in the deterioration in coding performance in conditions of efficiency and effectiveness. The problem is caused by the over-turned scope of vocabularies, which may take in unpredictable noises and make the precise terminology selection challenging. As a result, a corpus aware vocabulary terminology is naturally constructed by local mining approach, which can be used as terminology gap for further learning.

Let Q = {q1; q2;…..qn} and T = {t1; t2; ….tm} respectively represent a repository of QA pairs and their connected respectively denotes a repository of medical records and their connected locally mined terminologies. The target of global learning is to learn appropriate terminologies from the global vocabulary space T to interpret each medical record q in Q. Along with existing machine learning methods; graph-based learning achieves promising performance. In that paper, we also explore the graph-based learning model to accomplish 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 propose relationship identification and then we explain in detail, how to use our proposed model to associate the underlying connected medical records. Then, we present the optimal solution for learning model followed by the label bias estimation.

1) Relationship Identification:

The inter-terminology and inter-expert relationships are not intuitively implied from medical records, so we call them as implicit relationships. This subsection purpose is to introduce how to discover these kinds of relationships.

a) Inter-terminology Relationship: The medical terminologies in WordNet are organized into acyclic

taxonomic (is-a) hierarchies. For e.g., "viral pneumonia" is "infectious pneumonia" is-a "pneumonia" is-a "lung disease". Terminologies may have multiple parents. For e.g. "infectious pneumonia" is a child of "infectious disease". The inter-terminology hierarchical relationships are semantically capture well-defined ontology.

b) *Inter-expert Relationship* The inter-expert relationships will be viewed stronger if the experts are professionals in the same or related specific medical domain. This is returned by their historical data, i.e., the number of questions they have co-answered. Inspired by the Jaccard coefficient [27], the relationship between two experts ui and uj is calculated.

2) *Probabilistic Hypergraph Construction:*

The graph based learning models can be broadly categorized into simple graph-based and hypergraph based approaches, they are built on a graph where vertices are samples. The simple graph conveys the pair-wise relationship of vertices and overlooks the relations in higher orders, which are sensitive to the radius parameter used in similarity calculation [1]. As compared to simple graph, hypergraph contains the summarized local grouping information by allowing each hyperedge to connect more than two vertices simultaneously. Meanwhile hyperedge types and weights can be empirically set according to certain rules, and they can be heterogeneous to fuse comprehensive and diversified sources. Taken mutually, hypergraph-based learning partially fits task of terminology selection via integrating multi-faceted information cues, except considering the inter-terminology hierarchical relationship.

3) Global Learning Optimization:

Global learning optimization has been defined the hypergraph based framework for the global terminology learning that contains three objectives, and we aim to formulate each objective in details and derive a solution to this optimization problem. The ideas to formulate these three objectives are as follows. The first objective should guarantee that the relevance probability function is continuous and smooth in semantic gap. It means that the relevance probabilities of semantically similar medical records should be close to each other. The second objective is ensured by the empirical loss function, which forces the relevance probabilities to approach the initial roughly appropriate relevance scores. These two implicit constraints are widely adopted in reranking oriented methods. The third objective encourages the standards of medical records, which are connected by hierarchical structured terminologies, should be similar to each other.

4) Pseudo Label Estimation:

The idea of empirical loss term is to ensure the learnt relevance probabilities between terminologies and medical records are not far away from the initial roughly estimated relevance scores. Another method to decrease the size of the hypergraph is by pre-clustering the medical records during the data collection stage into several subgroups, and the hypergraph-based learning is conducted within each cluster. Each cluster contains the semantically close medical records, which are inter connected and most probably share the same vocabulary.

IV. Experimental Methodology and Results

This section explains the experimental evaluation of the proposed system. Firstly, we discuss the experimental settings, including the dataset and ground truth labelling. Dissimilar normal documents, these question samples are typically short, consisting of one or two sentences. They do not provide sufficient word co occurrences or shared contexts for effective similarity measure. It limits the accuracy obtained by the general learning approaches. In this work, we incorporate the answers to supplement the short questions, which will compensate for the data sparseness issue.

A. Dataset

We crawled more than 400 medical records from Health-Tap [24]. Each record contains question, answers, and all the involved experts who answered the question before. The questions with one answer or multiple answers but all from the same expert were eliminated, because they are isolated and unable to contribute much to the relationship analysis. The experts who replied less than four questions, along with the associated questions, were also removed. When we took a closer look at the dataset, and we found that it is reasonable: the non-active doctors generally do not concern their online reputation and seem not carefully to answer the questions within their own expertise.

B. Results

Time graph fig. 2 shows time required for retrieving questions answers in health care domain using local mining and local + global mining approaches. Proposed system takes less time to retrieve than existing approach.

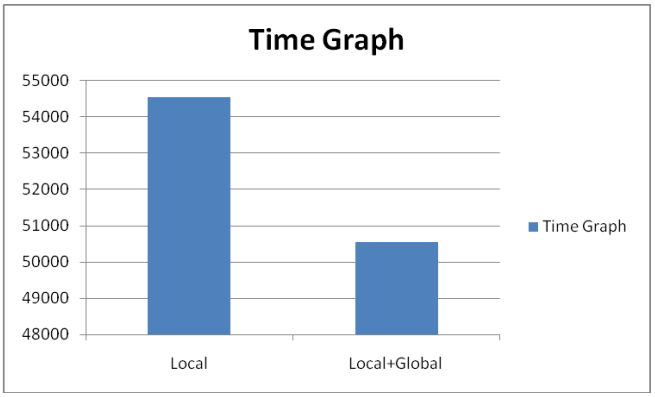

Fig .2 Time Graph

Table 1 shows the evaluation results for different question answer in terms of recall, precision and accuracy required to search that particular answers.

Using Local + Global Learning approach getting better results for bridging vocabulary gap between health seekers and health providers.

TABLE I. EVALUATATION RESULTS FOR LACAL AND GLOBAL MINING APPROACH

	Accuracy	Recall	Precision
Local	80	72	66
Local+Global	95	60	76

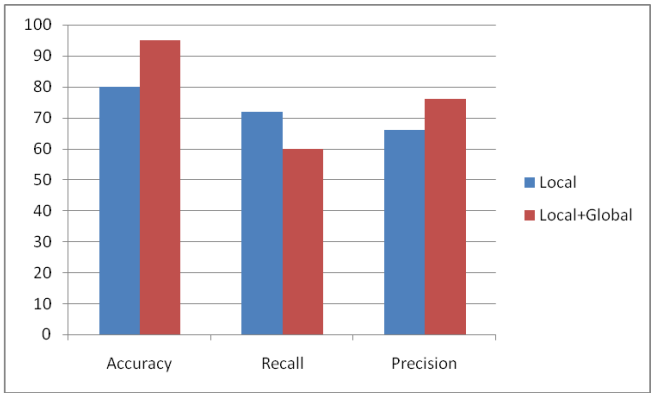

Fig .3 Evaluation Results for Local and Global Mining Approach

Fig. 3 shows the evaluation results for different question answers using local mining and local + mining approaches in terms of recall, precision and accuracy. From this, we conclude that local+ global mining approach gives good results than local mining approach.

V. CONCLUSION

This paper proposes a medical terminology assignment scheme to link the vocabulary gap between health seekers and healthcare knowledge. The scheme constitutes of local-global learning approach. Local mining establishes a tri-stage framework to locally code each medical record. The local mining approach, however, may suffer from information loss and low precision caused due to the absence of crucial medical concepts and the presence of the irrelevant medical concepts. It gives us motivation to propose a global learning approach to take care of the insufficiency of local coding approach. Global learning component collaboratively learns and propagates terminologies among underlying connected medical records. It implements the integration of heterogeneous information. Expanded evaluations on a real world dataset validate that our scheme is able to produce promising performance as compared to the prevailing coding methods. More importantly, the complete process of our approach is unsupervised and holds potential to handle large-scale data.

VI. FUTURE WORK

In future, we will investigate how to flexibly organize the unstructured medical content into user needs-aware ontology by leveraging the recommended medical terminologies.

ACKNOWLEDGMENT

We express our thanks to publishers, researchers for making their resource available & teachers for their guidance. We also thank the college authority for providing the required infrastructure and support. Last but not the least we would like to extend a heartfelt gratitude to friends and family members for their support.

REFERENCES

[1] Liqiang Nie, Yiliang Zhao, Mohammad Akbari, Jialie Shen, and T Chua. “Bridging the vocabulary gap between health seekers and healthcare knowledge”, In IEEE Transactions, 2014.

[2] Yan, Yan, et al. “Medical coding classification by leveraging inter code relationships”, of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[3] Fox, Susannah, and Maeve Duggan. “Health online 2013”, Health (2013).

[4] Pakhomov, Serguei VS, James D. Buntrock, and Christopher G. Chute. “Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques”, Journal of the American Medical Informatics Association 13.5 (2006): 516-525.

[5] Kim, Mi-Young, and Randy Goebel. “Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking”, Information Technology and Applications in Biomedicine (ITAB), 2010 10th IEEE International Conference on. IEEE, 2010.

[6] Larkey, Leah S., and W. Bruce Croft. “Automatic assignment of icd9 codes to discharge summaries”, University of Massachusetts (1995).

[7] Crammer, Koby, et al. “Automatic code assignment to medical text”, Proceedings of the Workshop on BioNLP 2007: Biological Translational, and Clinical Language Processing. Association for Computational Linguistics, 2007.

[8] Hersh, William R., and David H. Hickam. “Information retrieval in medicine: the SAPHIRE experience”, JASIS 46.10 (1995): 743-747

[9] Zou, Qinghua, et al. “IndexFinder: a method of extracting key concepts from clinical texts for indexing”, AMIA Annual Symposium Proceedings. Vol. 2003. American Medical Informatics Association, 2003.

[10] Leroy, Gondy, and Hsinchun Chen. “Meeting medical terminology needs-the ontology-enhanced Medical Concept Mapper”, Information Technology in Biomedicine, IEEE Transactions on 5.4 (2001): 261-270

[11] Lita, Lucian Vlad, et al. “Large Scale Diagnostic Code Classification for Medical Patient Records”, IJCNLP. 2008.

[12] Patrick, Jon, Yefeng Wang, and Peter Budd. “An automated system for conversion of clinical notes into SNOMED clinical terminology”, Proceedings of the fifth Australasian symposium on ACSW frontiers-Volume 68. Australian Computer Society, Inc., 2007.

[13] Dozier, Christopher, et al. “Fast tagging of medical terms in legal text”, Proceedings of the 11th international conference on Artificial intelligence and law. ACM, 2007.

[14] Nie, Liqiang, et al “Oracle in image search: A content-based approach to performance prediction”, ACM Transactions on Information Systems (TOIS)30.2 (2012): 13.

[15] Nie, Liqiang, et al. "Multimedia answering: enriching text QA with media information", Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.

[16] Nie, Liqiang, et al. "Harvesting visual concepts for image search with complex queries", Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012.

[17] Nie, Liqiang, et al. "Learning to Recommend Descriptive Tags for Questions in Social Forums", ACM Transactions on Information Systems (TOIS) 32.1 (2014): 5.

[18] Suominen, Hanna, et al. "Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description", Proceedings of the ICML/UAI/COLT Workshop on Machine Learning for Health-Care Applications. 2008.

[19] Rewatkar, Ashwini, and M. A. Potey. "Survey of Question Answering Approaches in Health Care".

[20] Nie, Liqiang, et al. "Wenzher: Comprehensive vertical search for healthcare domain", Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014.

[21] Nie, Liqiang, et al. "Beyond text QA: Multimedia answer generation by harvesting Web information", Multimedia, IEEE Transactions on 15.2 (2013): 426-441.

[22] http://wordnet.princeton.edu

[23] http://www.nlm.nih.gov/research/umls/Snomed/snomedmain.html

[24] https://www.healthtap.com