ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 10, March 2016

CROSSING OVER THE LEXICON GAP BETWEEN HEALTH SEEKERS AND PROVIDERS

*M.Supriya
Vins Chiristian Women's College Of Engineering

ABSTRACT—The vocabulary gap between health seekers and providers has hindered the cross-system operability and the inter-user reusability. To bridge this gap, this paper presents a novel scheme to code the medical records by jointly utilizing local mining and global learning approaches, which are tightly linked and mutually reinforced. Local mining extracts medical concepts from the QA pair itself and maps them to authenticated terminologies. Global learning, works towards enhancing the local mining via collaboratively discovering missing key terminologies and keeping off the irrelevant terminologies by analyzing the social neighbors. The rise of digital technologies has transformed the doctor-patient relationships. In this new era of web, when people struggle with their health concerns, most of them usually explore the Internet to research the problem before and after they see their doctors. A dictionary of vocabulary is constructed as a by-product, which is used as the terminology space for global learning.

Index Terms—Healthcare, medical terminology assignment, global learning, local mining, question answering

I INTRODUCTION

Information technologies are transforming the ways healthcare services are delivered, from patients' passively embracing their doctors' orders to patients' actively seeking online information that concerns their health. This trend is further confirmed by a national survey conducted by the Pew Research Center1 in Jan 2013, where they reported that one in three American adults have gone online to figure out their medical conditions in the past 12 months from the report time.

To better cater to health seekers, a growing number of community-based healthcare services have turned up, including HealthTap, HaoDF and WebMD. They are disseminating personalized health knowledge and connecting patients with doctors worldwide via question answering [1], [2]. 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. For patients, these systems provide nearly instant and trusted answers especially for complex and sophisticated problems. Over times, a tremendous number of medical records have been accumulated in their repositories, and in most circumstances, users may directly locate good answers by searching from these record archives, rather than waiting for the experts' responses or browsing through a list of potentially relevant documents from the Web.

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 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 non-standardized terms. Recently, some sites have encouraged experts to annotate the medical records with medical concepts. However, the tags used often vary wildly and medical concepts may not be medical terminologies [3]. 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 [4]. Automatically coding the medical records with standardized terminologies is highly desired. It leads to a consistent interoperable way of indexing, storing and aggregating across specialties and sites. In addition, it facilitates the medical record retrieval via bridging the vocabulary gap between queries and 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.

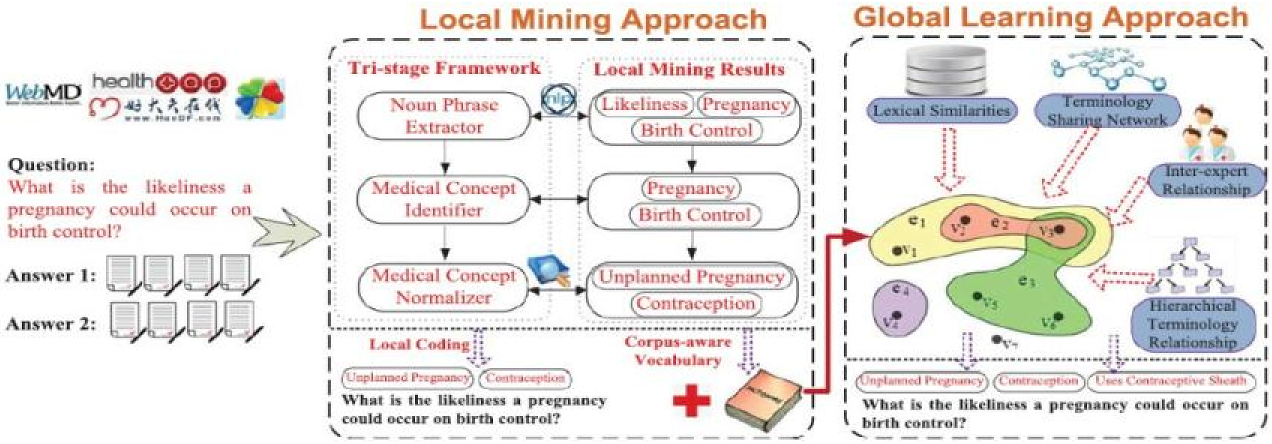

Fig.1.The schematic illustration of the proposed automatic medical terminology assignment scheme.

A novel scheme that is able to code the medical records with corpus-aware terminologies. As illustrated in Fig. 1, 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 maing 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 byproduct, a corpus-aware terminology vocabulary is naturally constructed, which can be used as terminology space for 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. 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 interexpert 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 respond to the same kinds of questionsprobably 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:

- The first work is 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 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.

II RELATED WORK

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. Back in 1995, Hersh and David 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 IndexFinder, 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 [9], [10], [11] 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 wellconstructed 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]. 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 evaluated on large-scale dataset and obtained promising performance [9]. Following that, a hierarchical model was studied in [5], 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. 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. 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. 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 examplebased rules and a machine learning module using Naïve Bayes. However, this integration is loosely coupled and the learning model can not incorporate heterogeneous cues, which is not a good choice for the community-based health services.

III LOCAL 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 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

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 interpretedas 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 [26], we assume that concepts that are relevant to medical domain occur frequently in medical domain and rarely in non-medical ones. Based on this assumption, we employ the concept entropy impurity [26] to comparatively measure the domain-relevance of a concept. For a concept c, its CEI is computed as follows:

$$CEI(c) = -\sum P(D_i|c) \log P(D_i|c)$$

where D1 and D2 respectively represents our medical corpus and a general-domain corpus; and $P(D_i|c)$ denotes the probability that a concept c is related to a specified domain Di; $P(D_i|c)$ can be computed as

$$P(D_i|c) = count(c, D_i)/count(c)$$

To remove the effect of different corpus's length, we define the normalized $P_n(D_i|c)$ as follows:

$$P_n(D_i|c) = (P_n(D_i|c)/L_i)/ (\sum P_n(D_j|c)L_j$$

where Li is the sum of document lengths in Di. Obviously, CEI(c) reaches the maximum value of 0:693, when concept c equally distributes within these two corpus. This implies that the larger CEI of a concept is, the more domain-irrelevant is it. To make it easily computer- processable, we define specificity of a concept to the medical domain as follows:

$Specificity(c) = 1 - \alpha\, CEI(c)$, if $P_n(D_1|c\ ) > P_n(D_2|c)$ otherwise $Specificity(c) = \alpha\, CEI(c)$

Where $\alpha = 0.5\ /0.693$ and a threshold is set to detect medical concepts.

C. Medical Concept Normalization

Although 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. Currently, there exist numerous authenticated vocabularies, including ICD,7 UMLS, and SNOMED CT.8 These medical and clinical terminologies were created in different times by different associations for different purposes. Take ICD as an example: it is typically used for external reporting requirements or other uses where data aggregation is advantageous. 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. The terminologies and their descriptions in SNOMED CT are first indexed.9 We then search each medical concept against the indexed SNOMED CT. For the medical concepts with multiple matched results, e.g., two results returned for "female", we keep all the returned terminology candidates (i.e., fully specified concept) for further selection. Concepts with the same or similar meanings in a natural language sense tend to be "close" in units of Google distance, while concepts with dissimilar meanings tend to be farther apart, we estimate the semantic similarity between the medical concept and the returned terminology candidates via exploring their cooccurrence on Google:

$$d(t_i,c) = \frac{\max(\log\ r(t_i\ ), \log r(c)) - \log r(t_i,c)}{\log G - \min(\log r(t_i\ ), \log r(c))}$$

where G is the total number of documents retrieved from Google; ti and c respectively represents the terminology candidate and the medical concept; r(x) is the number of hits for search concepts x; and $r(t_i,c)$is the number of web documents in which both ti and c co-occur. Then their semantic relevance is defined as:

$$S(t_i,c) = \exp\ (-\ d(t_i,c))$$

We then select the most relevant terminology candidate as the normalized result.

MEDICAL CONCEPT	NORMALIZED TERMINOLOGIES
Birth control	Contraception
Heart attack	Myocardial disorder
Breast cancer	Malignant tumor of breast
Blood loss	Hemorrhage
Home pregnancy test	Pregnancy test finding

Table1-Representative Medical Concepts after Normalization

SNOMED CT vocabulary is used as it provides the core general terminologies for the electronic health record and formal logic-based hierarchical structure. For example "birth control" is mapped with "contraception". It is essential to normalize is the key to bridge the vocabulary gap. The medical terminology hierarchy will enhance the 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 medical concept against the indexed SNOMED CT.

IV. GLOBAL LEARNING

Global learning is to complement the local medical coding in a graph-based approach. It collaboratively collects missing key concepts and propagates precise terminologies among underlying connected records over a large collection. The graph- based learning model is used to accomplish terminology selection task and 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. Christo Ananth et al. [13] proposed a method in which the minimization is per-formed in a sequential manner by the fusion move algorithm that uses the QPBO min-cut algorithm. Multi-shape GCs are proven to be more beneficial than single-shape GCs. Hence, the segmentation methods are validated by calculating statistical measures. The false positive (FP) is reduced and sensitivity and specificity improved by multiple MTANN.

They are both built on a graph where vertices are samples, while the simple graph conveys the pair-wise relationship of vertices and overlooks the relations in higher orders.

V. QA ARCHITECTURE

The main component of QA system summarized in following steps:

A. Question Analysis

The Question Analysis performance consists of classifying and analyzing the questions asked in natural language by the users can ask. This computational process is based on question classification. Question classification is assigning one of the generic patterns to each one of the questions that the user asks the system. This task starts once the user enters the question into the system. Question Analysis task firstly captures the semantics of the users question. These medical answer types can be diseases, symptoms, dose of drugs, and so on, according to the possible answers to the generic questions treated by the system. These question keywords are directly recognized by applying a set of heuristics to the predicates and the relationships between predicates in the logic form. Like question keywords QA system identifies complex nominal's and nouns recognized as medical expressions using Medical Named Entities Recognition including their possible adjective modifiers, the rest of the complex nominal's and nouns including their possible adjective modifiers and the main verb in the logic form. The logic form of a sentence is derived through applying NLP rules to the dependency relationship of the words in the sentence.

B. Document Retrieval

The document retrieval module can retrieve locally stored documents; its remote facility retrieves the relevant documents from medical websites using the Google search service. These medical websites are sorted from the previously defined medical website classification. This medical website classification is performed before the real-time execution of the google search engine and consists of defining the different medical website classes where the system can retrieve the medical documents. Document retrieval engine can start retrieving those relevant documents from medical websites whether there exists or not the association between the searched generic question and the medical websites. When the treated generic question has been related to at least one medical websites class then the Google search engine retrieves the relevant documents according to the question keywords in these medical websites.

C. Relevant Passage Selection

Relevant Passage Selection process consists of extracting the sentences from these medical documents that could answer questions of the user easily. These sentences are extracted by applying a technique based on comparing the question keywords in the documents and, those sentences that at least contain a question keyword are extracted from the document and are evaluated by the next Answer Extraction module that decides if the sentence correctly answers the user question.

D. Answer Extraction

Answer to any question asked by the user is extracted with the help of answer extraction module which extracts the answer by analyzing the sentences extracted by the previous relevant passage selection module. This module is performed by applying the following steps: the first one consists of inferring the logic form of the sentence and identifying the main verb in this logic form; the following step is to verify if this main verb belongs to the set of verbs that can answer the generic question; the third step is the acknowledgment of the medical entities in the logic form; the next step is of comparing if the medical entities searched as the answer is found in the logic form; and finally, the last step is the analysis of the predicates that relate the answer of the candidate, the main verb and the rest of the medical entities in the answer form. This module produces Ranking of Answers. The verb can distinctively relate two medical entities considering this feature as a direct link.

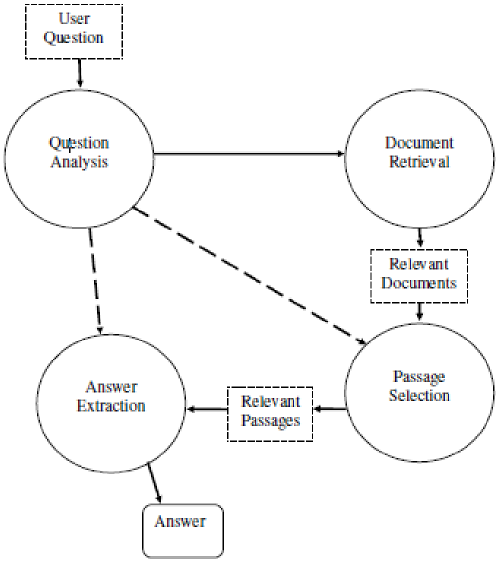

Fig 2: Medical QA System Modulate Architecture

VII CONCLUSIONS

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.

REFERENCES

[1] L. Nie, M. Akbari, T. Li, and T.-S. Chua, “A joint local-global approach for medical terminology assignment,” in Proc. Int. ACM SIGIR Conf., 2014.

[2] L. Nie, T. Li, M. Akbari, and T.-S. Chua, “Wenzher: Comprehensive vertical search for healthcare domain,” in Proc. Int. ACM SIGIR Conf., 2014, pp. 1245–1246.

[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] 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.

[5] L. Yves A., S. Lyudmila, and F. Carol, “Automating ICD-9-cm encoding using medical language processing: A feasibility study,” in Proc. AMIA Annu. Symp., 2000, p. 1072.

[6] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. Guo, “Fast tagging of medical terms in legal text,” in Proc. Int. Conf. Artif. Intell. Law, 2007, pp. 253–260.

[7] W. R. Hersh and H. David, “Information retrieval in medicine: The saphire experience,” J. Amer. Soc. Inf. Sci., vol. 46, no. 10, pp. 743–747, 1995.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] R. L. Cilibrasi and P. M. B. Vitanyi, “The google similarity distance,” IEEE Trans. Knowl. Data Eng., vol. 19, no. 3, pp. 370–383, Mar. 2007.

[13] Christo Ananth, G.Gayathri, M.Majitha Barvin, N.Juki Parsana, M.Parvin Banu, “Image Segmentation by Multi-shape GC-OAAM”, American Journal of Sustainable Cities and Society (AJSCS), Vol. 1, Issue 3, January 2014, pp 274-280

[14] J. Yu, D. Tao, and M. Wang, “Adaptive hypergraph learning and its application in image classification,”

IEEE Trans. Image Process., vol. 21, no. 7, pp. 3262–3272, Jul. 2012.

[15] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua, “Beyond text QA: Multimedia answer generation by harvesting webinformation,” IEEE Trans. Multimedia, vol. 15, no. 2, pp. 426–441, Feb. 2013.