

A Joint Local-Global Approach for Medical Terminology Assignment

Liqiang Nie
National University of
Singapore
nieliqiang@gmail.com

Mohammad Akbari
National University of
Singapore
akbari@nus.edu.sg

Tao Li
Zhejiang University
coylee917@gmail.com

Tat-Seng Chua
National University of
Singapore
chuats@nus.edu.sg

ABSTRACT

In community-based health services, vocabulary gap between health seekers and community generated knowledge has hindered data access. To bridge this gap, this paper presents a scheme to label question answer(QA) pairs by jointly utilizing local mining and global learning approaches. Local mining attempts to label individual QA pair by independently extracting medical concepts from the QA pair itself and mapping them to authenticated terminologies. However, it may suffer from information loss and lower precision, which are caused by the absence of key medical concepts and presence of irrelevant medical concepts. Global learning, on the other hand, works towards enhancing the local mining via collaboratively discovering missing key terminologies and keeping off the irrelevant terminologies by analyzing the social neighbors. Practically, this unsupervised scheme holds potential to large-scale data.

Categories and Subject Descriptors

J.3 [Life and Medical Sciences]: Health

Keywords

Community-based Health Services, Question Answers, Vocabulary Gap, Medical Terminology Assignment

1. BACKGROUND

The rise of digital technologies has transformed the patient-doctor relationships. Nowadays, when patients struggle with their health concerns, the majority usually explore the Internet to research the problem before and after they see their doctors. For example, 70% of Canadians turned to Internet to look up health-related information in 2009 [8] and 72% of American Internet users searched for

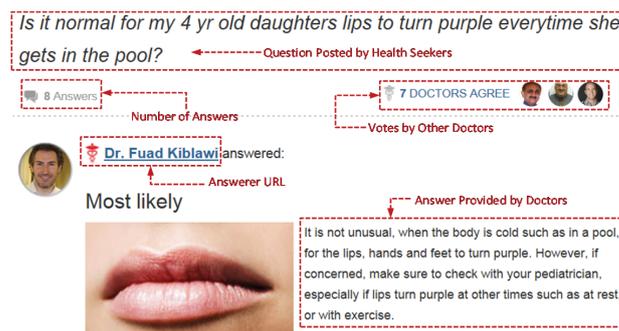


Figure 1: The illustration of a QA example from community-based health services (HealthTap).

health solutions in 2012 [4]. These metrics have reflected the scope and scale of the online health seekers.

To better serve the needs of health seekers, community-based health services have emerged as effective platforms for health knowledge dissemination and exchange, such as HealthTap¹, HaoDF² and WenZher[11]. They not only permit health seekers to freely post health-oriented questions, but also encourage doctors to provide trustworthy answers. Figure 1 demonstrates one typical QA pair example. Over time, a tremendous number of QA pairs has been accumulated in their repositories, and in most circumstances, health seekers may directly locate good answers by searching from these archives, rather than waiting for the experts' responses or painfully browsing through a list of documents from the general search engines.

2. CHALLENGES

In many cases, the community generated health content may not be directly usable due to the vocabulary gap, since participants with diverse backgrounds do not necessarily share the same vocabulary. Take HealthTap as an example. The same question may be described in substantially different ways by two individual health seekers. On the other hand, the answers provided by doctors may contain acronyms with multiple possible meanings, and non-standardized terms.

¹<https://www.healthtap.com/>

²www.haodf.com

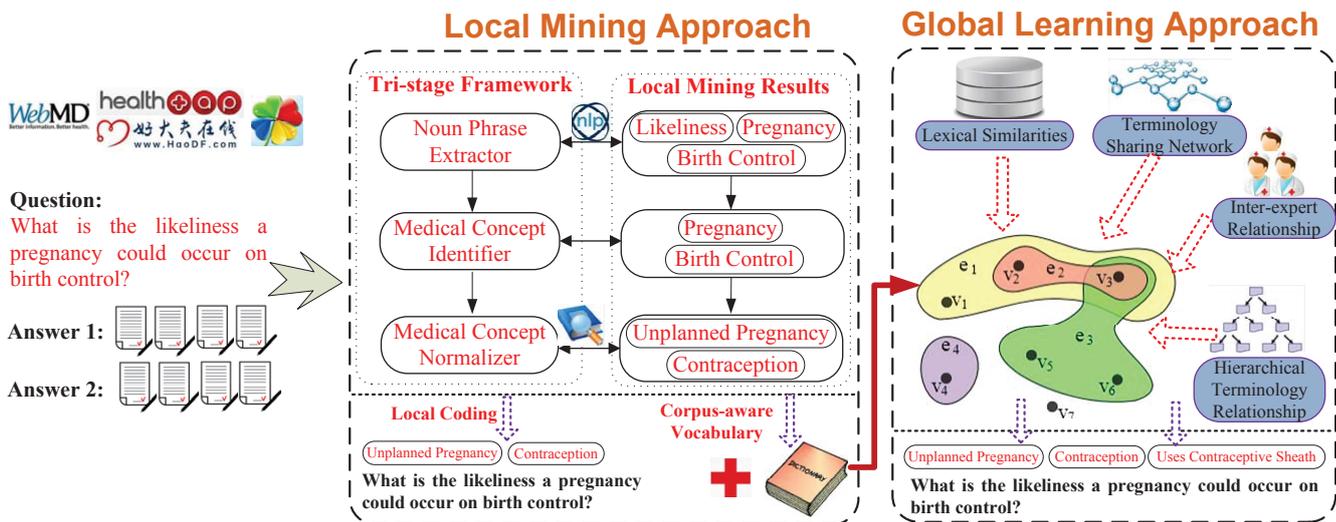


Figure 2: The schematic illustration of the proposed automatic medical terminology assignment scheme. The answer part is not displayed due to the space limitation.

In this work, we define medical concepts as medical domain-specific noun phrases, and medical terminologies 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. Even though some health communities have recently suggested doctors to annotate their answers with medical concepts, we cannot ensure that they are medical terminologies. Meanwhile, the tags adopted by doctors often vary greatly [3]. For example, “heart attack” and “myocardial disorder” are employed by different doctors to refer to the same medical diagnosis. It was shown that the inconsistency of community generated health data greatly hindered the cross-resource data exchange, management and integrity [9]. 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 [21]. Therefore, automatic coding of the QA pairs 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 QA pair retrieval via bridging the vocabulary gap between the queries and archives by coding the new queries with the standardized terminologies.

It is worth mentioning that there already exist several efforts dedicated to research on automatically mapping medical records to terminologies [19, 2, 10, 7, 17]. 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 this kind 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 external corpus independent knowledge may potentially 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.

3. METHOD

To overcome these limitations, we propose a novel scheme that is able to code the QA pairs with corpus-aware terminologies. As illustrated in Figure 2, the proposed scheme consists of two mutually reinforced components, namely, local mining and global learning.

3.1 Local Mining

Local mining aims to locally code the QA pairs by extracting the medical concepts from individual instance and then mapping them to terminologies based on the external authenticated vocabularies. To accomplish this task, we establish a tri-stage framework, which includes noun phrase extraction, medical concept detection and medical concept normalization.

To extract all the noun phrases, we initially assign part-of-speech tags to each word in the given QA pair by Stanford POS tagger³. We then extract tag sequences that match a fixed pattern of part-of-speech tags as noun phrases from the texts. This pattern is formulated as follows.

$$\begin{aligned} & (Adjective|Noun)^*(Noun\ Preposition) \\ & ?(Adjective|Noun)^*Noun. \end{aligned} \quad (1)$$

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”.

Inspired by the efforts in [18, 6], in order to differentiate the medical concepts from other general noun phrases, we assume that concepts that are relevant to medical domain occur frequently in medical domain and rarely in

³<http://nlp.stanford.edu/software/tagger.shtml>

non-medical ones. Based on this assumption, we employ the concept entropy impurity (CEI) [6] to comparatively measure the domain-relevance of a concept by comparing the term frequencies between two different corpora D_1 and D_2 . D_1 is our medical-domain corpus and D_2 is a general English Gigaword data of Linguistic Data Consortium⁴.

As aforementioned, we cannot ensure that all medical concepts 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 an appropriate external standardized dictionary and this normalization is the key to bridging the vocabulary gap. In this work, we use SNOMED CT⁵ as our dictionary, since 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 indexed⁶. 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 for further selection. Enlightened by Google distance [1], we estimate the semantic similarity between the medical concept and the returned terminology candidates via exploring their co-occurrence on Google. We then select the most relevant terminology candidate as the normalized result.

Local mining, however, may suffer from various problems. The first problem is incompleteness. This is because some key medical concepts may not explicitly present in the QA pairs. The QA pair illustrated in Figure 2 shows an example of this situation, where the accurate terminology: “use contraceptive sheath” is absent from the QA pair. The second one is the lower precision. This is due to some irrelevant medical concepts explicitly embedded in the QA pairs, and are mistakenly detected and normalized by the local approach. For instance, given the question, “*What are the risks getting pregnant and giving birth later in life ?*”, the terminology “finding of life event” as normalized from the irrelevant medical concept “life” is assigned as code. It is less informative to capture the main intent.

3.2 Global Learning

It is noteworthy that most previous efforts, including our local approach, attempted to map the QA pairs directly to the entries in external dictionaries without any pruning. This approach often presents problems since the external dictionaries usually cover relatively comprehensive terminologies and are far beyond the vocabulary scope of the given corpus. It may result in the deterioration in coding performance in terms of efficiency and effectiveness. The problem is caused by the over-widened scope of vocabularies, which may bring in unpredictable noises and make the precise terminology selection challenging. As a byproduct, a corpus-aware terminology vocabulary is naturally constructed by our local mining approach, which can be used as terminology space for further learning.

Let $\mathcal{Q} = \{q_1, q_2, \dots, q_N\}$ and $\mathcal{T} = \{t_1, t_2, \dots, t_M\}$ respectively denote a repository of QA pairs and their associated

locally mined terminologies. The target of global learning is to learn appropriate terminologies from the global terminology space \mathcal{T} to annotate each q in \mathcal{Q} . In this work, the global learning task is regarded as a multi-label learning problem[16]. It is formulated as,

$$\arg \min_{\mathbf{F}} \sum_{i=1}^M \left\{ \Omega(\mathbf{f}_i) + \lambda L(\mathbf{f}_i) + \mu \sum_{j=1}^M R_{ij} \|\mathbf{f}_i - \mathbf{f}_j\|^2 \right\}, \quad (2)$$

where M refers to the number of classes, i.e., the number of medical terminologies to be assigned. Vector \mathbf{f}_i is the i th column of \mathbf{F} , representing the relevance scores of each QA pair to the i -th terminology. $\Omega(\mathbf{f})$ and $L(\mathbf{f})$ denotes the regularizer on the hypergraph and empirical loss, respectively. In addition, R_{ij} is the inter-terminology relationship between terminology i and terminology j . They are mined by exploiting the external well-structured ontology, which are able to alleviate the granularity mismatch problems and reduce the irrelevant sibling terminologies. By differentiating the above equation with respect to \mathbf{F} , we can obtain a closed-form solution.

The philosophy to formulate these three objectives is as follows. The first objective aims to guarantee that the relevance probability function is continuous and smooth in semantic space. This means that the relevance probabilities of semantically similar QA pairs 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 estimated relevance scores. These two implicit constraints are widely adopted in reranking-oriented approaches [12, 13, 14, 15]. The last encourages the values of QA pairs, which are connected by hierarchical structured terminologies, to be similar to each other.

When it comes to hypergraph construction, the N QA pairs from \mathcal{Q} are regarded as vertices and they are connected by three types of hyperedges. The first type takes each vertex as a centroid and forms a hyperedge by circling around its k -nearest neighbors based on QA pair content similarities. This procedure was first adopted in [5]. The second type is based on terminology-sharing network. For each terminology, it groups all the QA pairs sharing the same terminology together. The third type actually takes the users’ social behaviours into consideration by rounding up all the questions answered by closely associated doctors. The inter-doctor relationships are inferred from the doctors’ historical data. Specifically, doctors who are frequently respond 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. As a consequence, up to $N + M + U$ hyperedges are constructed in our hypergraph, where U is the number of involved doctors. Learning from this hypergraph, we are able to find missing key concepts and propagate precise terminologies among underlying connected records over a large collection. Besides the semantic similarity among QA pairs and terminology-sharing network, the inter-terminology and inter-expert relationships are seamlessly integrated in the proposed model. It is noteworthy that a rich set of healthcare specific features are extracted and weighted for similarity estimation.

4. EXPERIMENTS

We crawled more than 109 thousand QA pairs from

⁴<http://www ldc upenn edu/>

⁵<http://www ihtsdo org/snomed-ct/>

⁶<http://viw2 vetmed vt edu/sct/menu cfm>

Table 1: The comparative evaluation results of medical terminology assignment in terms of $S@K$ and $P@K$.

Approach \ Metric	S@1	S@2	S@3	S@4	P@1	P@2	P@3	P@4
LocalMining	72.0%	84.0%	91.0%	95.0%	72.0%	72.1%	69.7%	68.3%
Local+Global	83.0%	92.0%	98.0%	100.0%	83.0%	81.5%	80.3%	78.8%

Table 2: Comparative illustration of the representative question samples with locally mined terminologies and locally+globally recommended terminologies. Answers are not displayed due to limited space.

QA pairs	Locally Mined Terminologies	Local Mining + Global Learning
Is it safe to color my hair during pregnancy ?	hair structure, dyed hair, feeling safe, patient currently pregnant, first trimester pregnancy...	hair structure, patient currently pregnant, coal tar allergy, hair color change, disorder of endocrine system...
If I get an infection caused by gum disease, can that be transferred to my fetus ?	infectious disease, gingival disease, entire fetus, inflammation, periodontal disease...	infectious disease, prematurity of fetus, gingival disease, periodontal disease low birth weight infant...

HealthTap, which involve 5,958 unique doctors. For ground truth construction, we invited three professionals with master degrees majored in medicine programme. The labelers were trained with a short tutorial and a set of demonstrating examples. A majority voting scheme among the three labelers can partially alleviate the subjectivity problem. The annotators were required to label only top five recommended terminologies for each QA pair, and they were labeled either as “positive” or “negative”. 100 QA pairs were labeled as testing set.

We adopted two metrics that are able to characterize precisions from different aspects. The first is average $S@K$ over all testing QA pairs, which measures the probability of finding a relevant terminology among the top K recommended ones. To be specific, for each testing QA pair, $S@K$ is assigned to 1 if a relevant terminology is positioned in the top K and 0 otherwise. The second one is average $P@K$ that stands for the proportion of recommended terminologies that are relevant[20]. $P@K$ is defined as $P@K = \frac{|\mathcal{C} \cap \mathcal{R}|}{|\mathcal{C}|}$

where \mathcal{C} is a set of the top K terminologies, and \mathcal{R} is the manually labeled positive ones.

Table 1 displays the comparison. We can see that the local mining approach achieves the worst performance. This is reasonable, because irrelevant concepts may be mapped to terminologies because of their presence in the QA pairs.

Table 2 comparatively illustrates the representative QA pair samples with locally minded terminologies and locally+globally recommended ones. Intuitively, the terminologies are more comprehensive and reliable after enhancement with global learning.

5. CONCLUSIONS AND FUTURE WORK

This paper presented a medical terminology assignment scheme to bridge the vocabulary gap between health seekers and community generated knowledge. A strong unified framework of local mining and global learning is proposed to tackle this research issue, instead of the conventional isolated utilization. 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. In addition, it builds a novel global learning model to enhance the local coding results. This model seamlessly integrates various heterogeneous cues.

In the future, we will investigate how to flexibly organize

the unstructured medical content into user needs-aware ontology by the recommended medical terminologies.

6. ACKNOWLEDGEMENTS

This work was supported by NUS-Tsinghua Extreme Search project under the grant number: R-252-300-001-490.

7. REFERENCES

- [1] R. L. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 2007.
- [2] C. Dozier, R. Kondadadi, K. Al-Kofahi, M. Chaudhary, and X. Guo. Fast tagging of medical terms in legal text. In *Proceedings of the International Conference on Artificial Intelligence and Law*, 2007.
- [3] A. e-HIM Work Group on Computer-Assisted Coding. Delving into computer-assisted coding. *Journal of American Health Information Management Association*, 2004.
- [4] S. Fox and M. Duggan. Health online 2013. Survey, Pew Research Center, 2013.
- [5] Y. Huang, Q. Liu, S. Zhang, and D. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.
- [6] M.-Y. Kim and R. Goebel. Detection and normalization of medical terms using domain-specific term frequency and adaptive ranking. In *Information Technology and Applications in Biomedicine, IEEE International Conference on*, 2010.
- [7] L. S. Larkey and W. B. Croft. Automatic assignment of icd9 codes to discharge summaries. *PhD Thesis, University of Massachusetts at Amherst*, 1995.
- [8] M. Law. Online drug information in canada. Technical report, 2012.
- [9] G. Leroy and H. Chen. Meeting medical terminology needs-the ontology-enhanced medical concept mapper. *IEEE Transactions on Information Technology in Biomedicine*, 2001.
- [10] L. V. Lita, S. Yu, S. Niculescu, and J. Bi. Large scale diagnostic code classification for medical patient records. In *Proceedings of the Conference on Artificial Intelligence in Medicine*, 1995.
- [11] L. Nie, T. Li, M. Akbari, J. Shen, and T.-S. Chua. Wenzher: Comprehensive vertical search for healthcare domain. In *Proceedings of the International ACM SIGIR Conference*, 2014.
- [12] L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua. Beyond text qa: Multimedia answer generation by harvesting web information. *IEEE Transactions on Multimedia*, 2013.
- [13] L. Nie, M. Wang, Z.-J. Zha, and T.-S. Chua. Oracle in image search: A content-based approach to performance prediction. *ACM Transactions on Information System*, 2012.
- [14] L. Nie, M. Wang, Z.-J. Zha, G. Li, and T.-S. Chua. Multimedia answering: Enriching text qa with media information. In *Proceedings of the International ACM SIGIR Conference*, 2011.
- [15] L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua. Harvesting visual concepts for image search with complex queries. In *Proceedings of the International Conference on Multimedia*, 2012.
- [16] L. Nie, Y.-L. Zhao, X. Wang, J. Shen, and T.-S. Chua. Learning to recommend descriptive tags for questions in social forums. *ACM Transactions on Information System*, 2014.
- [17] H. Suominen, F. Ginter, S. Pyysalo, A. Airola, T. Pahikkala, S. Salanterä, and T. Salakoski. Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. In *Proceedings of the ICML Workshop on Machine Learning for Health-Care Applications*, 2008.
- [18] P. Velardi, M. Missikoff, and R. Basili. Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the workshop on Human Language Technology and Knowledge Management*, 2001.
- [19] L. Yves A., S. Lyudmila, and F. Carol. Automating icd-9-cm encoding using medical language processing: A feasibility study. In *Proceedings of the AMIA Annual Symposium*, 2000.
- [20] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua. Personalized recommendations of locally interesting venues to tourists via cross region community matching. *ACM Transactions on Intelligent Systems and Technology*, 2013.
- [21] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt. Exploiting medical hierarchies for concept-based information retrieval. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, 2012.