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.

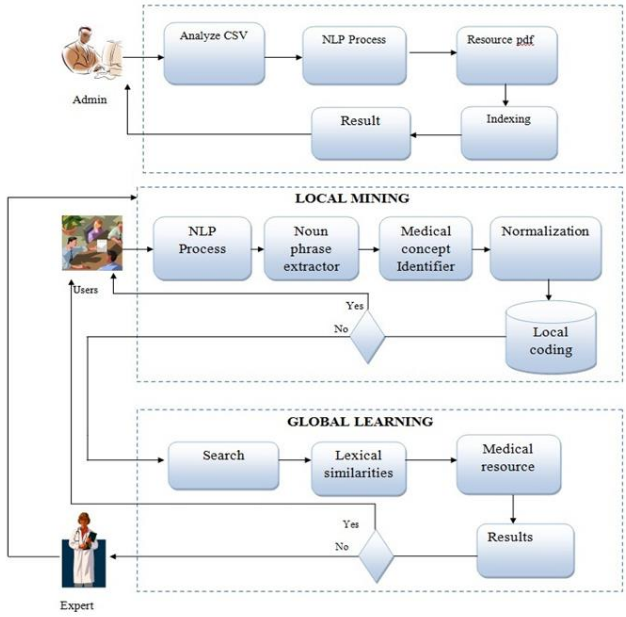

Figure 1: Architecture Diagram

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. 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 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, 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 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 locally, which naturally construct a corpus-aware terminology vocabulary with the help of external knowledge. Fig. 1. The schematic illustration of the proposed automatic medical terminology assignment scheme. The answer part is not displayed due to the space limitation.

NIE ET AL.: BRIDGING THE VOCABULARY GAP BETWEEN HEALTH SEEKERS AND HEALTHCARE KNOWLEDGE 397

It builds a novel global learning model to collaboratively enhance the local coding results. This model seamlessly integrates various heterogeneous information cues.

The remainders are structured as follows. Section 2 briefly reviews the related work. The local mining and global learning approaches are respectively introduced in Sections 3 and 4. Section 5 details the experimental results and analysis, followed by our concluding remarks in Section 7.

II. METHODS AND MATERIAL

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 evaluated 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 Na€ıve Bayes. However, this integration is loosely coupled and the learning model cannot incorporate heterogeneous cues, which is not a good choice for the community-based health services.

LOCAL MINING

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 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.

Q and A Application

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

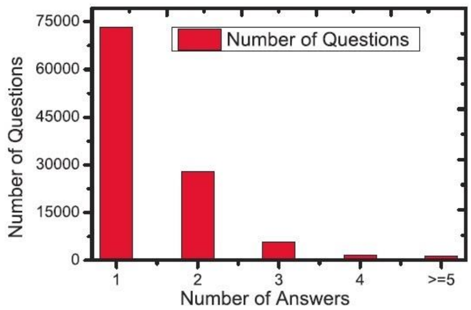

Figure 2: distribution of questions with respect to their received answers.

NLP Process

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

III. RESULTS AND DISCUSSION

Bridging Gap

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

Machine Learning

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

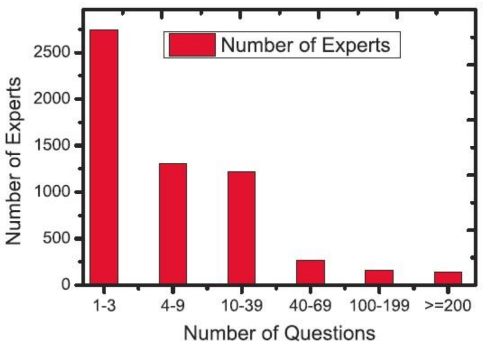

Fig. 3. The distribution of experts with respect to the number of questions they answered.

IV. CONCLUSION

This paper presents a medical terminology assignment scheme to provide the instant answer for the health seeker using machine learning. The instant answers are provided by manning and experts who answers for the queries post by the user which has a major advantage of time consistency and exact answers. In the future enhancement we are trying to develop with the queries by uploading images which is more convenient for experts to answer.

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

[4] 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.

[5] 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.

[6] E. J. M. Laur_ıa 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.

[7] 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.

[8] 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.

[9] L. V. Lita, S. Yu, S. Niculescu, and J. Bi, "Large scale diagnostic code classification for medical patient records," in Proc. Conf. Artif. Intell. Med., 1995.

[10] L. S. Larkey and W. B. Croft, "Automatic assignment of icd9 codes to discharge summaries," PhD dissertation, Dept. Comput. Sci., Univ. Massachusetts at Amherst, Amherst, MA, USA, 1995.

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

[12] J. Patrick, Y. Wang, and P. Budd, "An automated system for conversion of clinical notes into snomed clinical terminology," in Proc. 5th Australasian Symp. ACSW Frontiers, 2007, pp. 219–226.

[13] 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.

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

[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.